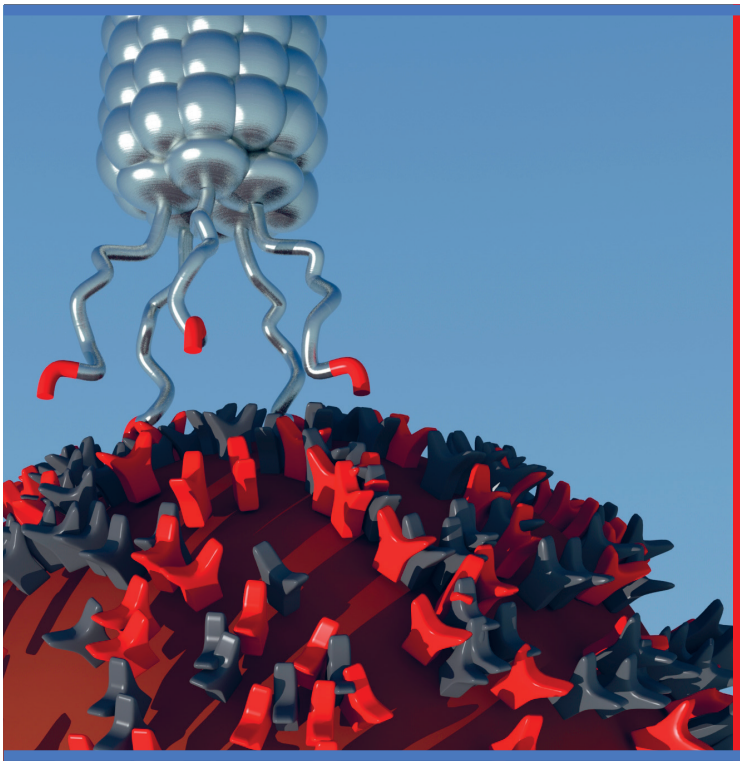
High-Throughput Tools for Characterization of Antibody Epitopes

Christiansen, Anders

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record



High-Throughput Tools for Characterization of Antibody Epitopes

Anders Christiansen
PhD Thesis August 2015

DTU Nanotech
Department of Micro- and Nanotechnology

DTU Nanotech
Department of Micro- and Nanotechnology

High-Throughput Tools for Characterization of Antibody Epitopes

PhD dissertation

Anders Christiansen

DTU Nanotech

The Technical University of Denmark

2015

Preface

This PhD dissertation covers the research that I have been involved in at The Department of Micro- and Nanotechnology (DTU Nanotech) at The Technical University of Denmark (DTU), Kgs. Lyngby, Denmark. The research was carried out under the supervision of Martin Dufva and Katrine L Bøgh in a period from September 2012 to August 2015. The work has resulted in four studies presented in this dissertation, including two original manuscripts:

1. "High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum"
Christiansen A, Kringelum JV, Hansen CS, Bøgh KL, Sullivan E, Patel J, Rigby NM, Eiwegger T, Szépfalusi Z, Masi FD ,Nielsen M, Lund O, Dufva M. Scientific Reports 2015 Aug 6;5:12913. doi: 10.1038/srep12913.

2. "High-resolution epitope mapping of peanut allergens demonstrates individualized and highly persistent antibody binding patterns at the residue level"
Hansen CS, Dufva M, Bøgh KL, Sullivan E, Patel J, Eiwegger T, Szépfalusi Z, Nielsen M, Christiansen A
Manuscript in preparation

In addition, the research has been disseminated at the following conferences:

1. International Symposium on Molecular Allergology (ISMA) Vienna, Austria. 5-7 December 2013.
The conference abstract was published in Clinical and Translational Allergy, vol 4 , no. Suppl 2, pp. P27., 10.1186/2045-7022-4-S2-P27.
2. The Gordon Research Conference – Antibody Biology & Engineering, Lucca, Italy. 22-28 March 2014.
Oral presentation and poster presentation.
3. The European Academy for Allergy and Clinical Immunology (EAACI) Congress 2014, Copenhagen, Denmark. 7-11 June 2014.
Poster discussion session. Session prize-winner. Conference abstract in ALLERGY, Vol. 69, No. Supplement, 297, 2014, p. 140-140.
4. The European Academy for Allergy and Clinical Immunology (EAACI) Congress 2015, Barcelona, Spain. 6-10 June 2014.
Oral presentation and poster presentation. Conference abstract accepted for publication in ALLERGY.

Table of Contents

Acknowledgements

First and foremost I would like to thank my supervisor Martin Dufva and DTU Nanotech for giving me the opportunity to conduct my PhD studies at DTU. I would like to thank Martin for countless scientific discussions, great input to my work, fun brainstorming sessions and for encouraging me to follow my own ideas. I would also like to express my gratitude towards my co-supervisor, Katrine L Bøgh, at DTU Food and thank her for introducing me to the allergy field and for her detailed input on many projects. I have also appreciated her company at various conferences around Europe. At DTU Nanotech, I would like to thank the FAST group and my colleagues in building 423 for creating a pleasant atmosphere to work in.

I greatly benefitted from a close collaboration that was established with the Center for Biological Sequence Analysis (CBS) at DTU. In particular, I would like to express my gratitude to Christian Skjødt and Jens Kringelum for an extensive collaborative effort resulting in the first manuscript. I appreciate the introduction to bioinformatics, UNIX as well as countless discussions on large and small decisions. I would also like to thank Christian for his continued effort and collaboration on the second manuscript. Together we have solved many obstacles, big and small. I also appreciate the input and guidance from Ole Lund and Morten Nielsen on different projects. I am also very grateful for the assistance with deep sequencing provided by DMAC, especially the effort by Neslihan Bicen in solving various issues during the process.

I am very grateful that Thomas Eiwegger and Zsolt Szépfalusi at the Medical University of Vienna, Austria accepted to be part of these studies and provided patient samples as well as their clinical perspective to discussions on allergy. I would also like to acknowledge the contributions of the patients whose samples were used in these studies and express my appreciation that they accepted to contribute to allergy research. I am grateful to Neil Rigby at the Institute of Food Research, United Kingdom for putting in the effort to provide purified peanut allergens.

I am also very grateful to Jigar Patel and Eric Sullivan at Roche NimbleGen for allowing the access to a peptide microarray platform that has yet to become commercially available. I also appreciate their experience and input on experimental design and trouble-shooting regarding the arrays. I wish to thank Andreas Laustsen at Copenhagen University (KU) for our collaboration on the search for cobratoxin inhibitors. I would further like to thank Brian Lohse, Jónas Johanneson and Timothy Lynagh, all at KU, for

their contributions to the cobra project. I am also thankful for the collaboration with the people at Symphogen, especially Jenifer Clausell-Tormos and Nikolaj Dietrich, and all the other participants in the ARiM project.

The studies were financially supported by the Technical University of Denmark, DTU Nanotech, The Danish National Advanced Technology Foundation, Dr. Techn. A.N. Neergaard and Wife's Foundation and the Oticon Foundation.

Finally, I would like to thank my family, especially my parents for their continued support and for contributing greatly to who I am and where I am today. I am very grateful to my father for teaching me the curiosity and patience that is needed to do research and my mother for inspiring me to always be helpful and supportive of the people around me. My deepest gratitude also goes to my wonderful girlfriend, Birgitte, who fills my life with cheerful moments and support me when I need it.

Anders Christiansen

Copenhagen, August 2015

Abbreviations

BLAST: Basic local alignment search tool

BSA: Bovine serum albumin

CDR: Complementarity determining region

CDRH: Complementarity determining region heavy chain

DMAC: DTU multi-assay core facility

ELISA: Enzyme-linked immunosorbent assay

Fab: Fragment antibody binding

HIV: Human immunodeficiency virus

HRP: Horseradish peroxidase

HTS: High-throughput sequencing

Ig: Immunoglobulin

Indel: Insertion-deletion

KU: Copenhagen University

nAChR: Nicotinic acetylcholine Receptor

NMR: Nuclear magnetic resonance

OIT: Oral immunotherapy

OPD: *ortho*-Phenylenediamine

ORF: Open reading frame

PBS: Phosphate-buffered saline

PEG: Poly-ethylene glycol

RT: Room temperature

scFv: Single-chain variable fragment

SMP: Skim milk powder

TUP: Target-unrelated peptide

TEVC: two-electrode voltage clamp

Abstract

Antibodies are molecules of tremendous importance. In their primary role, they protect our bodies against disease. However, in recent decades, scientists have harnessed the binding capabilities of antibodies and have applied them widely in research, diagnostics and therapeutics. Consequently, it is important to characterize antibodies thoroughly. In parallel to the characterization of antibodies, it is also important to characterize the binding area that is recognized by the antibody, known as an epitope. With the development of new technologies, such as high-throughput sequencing (HTS), it is important to determine how these methods can improve our understanding of antibodies and their epitopes. The overall objective of the presented studies was to investigate how emerging technologies (specifically HTS coupled with phage display and next-generation peptide microarrays) could be used for epitope mapping.

In **Chapter 1**, it was examined whether combining phage display, a traditional epitope mapping approach, with HTS would improve the method. The developed approach was successfully used to map Ara h 1 epitopes in sera from patients with peanut allergy. Notably, the sera represented difficult biological samples due to the rarity of the relevant antibodies and the polyclonal nature of serum. The inclusion of control samples enabled the development of a bioinformatic approach that identified peptide motifs of interest based on clustering and contrasting. A widespread problem in phage display, which is the unintended selection of peptides that are target-unspecific, was examined by comparing patient and control samples. The experiments highlighted that HTS can potentially improve on phage display by enabling the analysis of complex biological samples. Coupling the two methods furthermore has the capacity to omit traditional clone picking and functional testing which is a laborious part of phage display.

In the following study, **Chapter 2**, it was described how the approach developed in Chapter 1 could be utilized for a different application of phage display, specifically the identification of peptide binders. In this study, phage display screenings were used to identify peptides that could inhibit a major toxin in cobra snake venom, α-cobratoxin. Peptide inhibitors were successfully identified. Importantly, HTS enabled the identification of toxin inhibitors that were not discovered by traditional phage display.

Phage display coupled with HTS was again used in **Chapter 3** in an attempt to map the epitopes of a therapeutic target injected into animals. The animals were immunized with a therapeutic target and the expectation was that they develop antibodies, which can be used in therapy. While no epitopes could be

definitively identified, the study demonstrated the potential of the MiSeq HTS platform. Sequencing of the phage library also showed that many of the target-unrelated phages identified in the previous chapters, were frequent in the original library, thus indicating that they held proliferation advantages.

Finally, in **Chapter 4**, a different emerging technology, next-generation peptide microarrays, was applied for epitope mapping of major peanut allergens using sera from allergic patients. New developments in the peptide microarray have enabled a greatly increased throughput. In this study, these improvements were utilized to characterize epitopes at high resolution, i.e. determine the importance of each residue for antibody binding, for all major peanut allergens. Epitope reactivity among patients often converged on known epitope hotspots, however the binding patterns were somewhat heterogeneous when examined at the residue level. A high degree of correlation between IgE and IgG4 epitope binding patterns were observed, possibly indicating a common clonal origin. Finally, since the patients had been sampled over time it could be confirmed that the epitope binding patterns were stable over multiple years.

Taken together, the presented studies demonstrated new applications for the investigated techniques focusing on their utilization in epitope mapping. In the process, new insights were obtained into how antibodies recognize their targets in a major disease, i.e. food allergy.

Dansk resume

Antistoffer er enormt vigtige molekyler. Deres oprindelige rolle er at beskytte vores krop mod sygdomme. I de seneste årtier har forskere dog udnyttet antistoffers særlige bindings-egenskaber inden for forskning og udvikling af ny diagnostik og nye lægemidler. Derfor er det vigtigt at karakterisere antistoffer grundigt. Parallelt med karakteriseringen af antistofferne er det også vigtigt at karakterisere antistoffernes bindings-område kendt som epitoper. Med udviklingen af ny teknologi, f.eks. genom-sekventering, er det vigtigt at undersøge om denne teknologiudvikling kan forbedre vores forståelse af antistoffer og deres epitoper. Det overordnede mål med de præsenterede studier var derfor at undersøge hvordan denne teknologiudvikling (specifikt genom-sekventering kombineret med 'phage display' og den nye generation af peptid microarrays) kan udnyttes til at kortlægge epitoper.

I **Kapitel 1** blev det undersøgt hvorvidt den traditionelle 'phage display' metode kunne forbedres ved at koble den med genom-sekventerings teknologi. Den udviklede tilgang blev med succes brugt til at kortlægge Ara h 1 epitoper i sera fra patienter med jordnødde- (peanut-) allergi. Sera-prøverne repræsenterer vanskeligt biologisk materiale med sjældne antistoffer i en polyklonal blanding. Bioinformatisk programmer blev udviklet som kunne identificere peptid-motiver baseret på gruppering af ensartede peptider og sammenligning med kontrolprøver. En sammenligning af patient- og kontrolprøver bekræftede at uspecifikke peptider er et problem i 'phage display'. Forsøgene viste at genom-sekventering potentielt kan forbedre 'phage display' ved at reducere antallet af selektions runder og tillade analyse af komplekse biologiske prøver. Desuden kan man muligvis omgå den traditionelle isolering af enkelt-kloner og funktionelle test.

I det efterfølgende studie, præsenteret i **Kapitel 2**, blev det beskrevet hvordan den metode der blev udviklet i Kapitel 1 kunne anvendes til en anden hyppig 'phage display' applikation: identificeringen af peptid-bindere. I dette studie blev 'phage display' brugt til at finde peptider der kunne blokere et vigtigt slange-toxin (α-cobratoxin) fra cobraslanger. Det lykkedes at finde sådanne inhibitorer og genom-sekventering hjalp med at identificere peptider som blev overset ved traditionel 'phage display'.

Koblingen af 'phage display' og genom-sekventering blev igen brugt i **Kapitel 3** i et forsøg på at kortlægge epitoper fra et terapeutisk relevant protein. Et antal dyr blev immuniseret med det terapeutisk relevante protein i forventning om at dyrene udviklede antistoffer som kunne anvendes som lægemidler. Der kunne ikke definitivt identificeres relevante epitoper, men studiet demonstrerede potentialet for MiSeq sekventerings-platformen. Sekventering af phage biblioteket viste at mange af de

uspecifikke peptider, der var blevet identificeret i tidligere kapitler, var hyppige i det oprindelige bibliotek, hvilket indikerede at de havde fordele under opformeringen i bakterier.

I **Kapitel 4** blev den nye generation af peptid microarrays brugt til at kortlægge epitoper i sera fra patienter med jordnødde-allergi. Forbedringer af microarray platformen har øget mulighederne for hvad den kan anvendes til. I dette studie blev forbedringerne udnyttet til at karakterise epitoper i høj detaljegrad og bestemme bidraget fra hver aminosyre til bindingen af antistof. Det blev gjort for alle de vigtigste peanut allergener. Selvom patienter på det overordnede niveau så ud til at binde en epitop på en ensartet måde viste den store detaljegrad at bindingsmønstrene var mere heterogene, når de blev undersøgt på aminosyre niveau. Desuden var der en høj korrelation mellem IgE og IgG4 bindingsmønstre hvilket kunne indikere fælles klonal oprindelse. Det blev derudover demonstreret at epitop bindingsmønstrene var meget stabile over en årrække.

Samlet set demonstrerede studierne nye anvendelser for de undersøgte metoder med fokus på deres anvendelse inden for kortlægning af epitoper. Gennem processen blev ny viden om hvordan antistoffer genkender deres mål i fødevareallergi opnået.

Introduction

Antibodies – Tools of specificity in our immune system

We live our lives constantly surrounded by microorganisms. However, in spite of the fact that many of them are disease causing, we rarely become ill. This feat can be attributed to our immune system; its primary role is to defend us against infections.

The immune system is an intricate machinery with many gears and cogs. Yet, an essential tool is the antibody (also known as an immunoglobulin). The overarching role of antibodies is in distinguishing foreign elements, antigens, from components that are intrinsic to the body.

In general, antibodies work in three ways[1]: First of all they can bind to a pathogen thereby preventing its interaction with the target cell surface, a process known as neutralization. Secondly, they can coat the surface of a pathogen and enhance phagocytosis, known as opsonization. Finally, they can activate proteins of the complement system.

Antibodies have two core features that make them extraordinary tools in the immune system, their specific binding and huge diversity. Grooves and protrusions at the hypervariable region of the antibody are well suited for specific binding with high-affinity[1]. Several mechanisms have evolved to allow for an immensely varied antibody pool, with a theoretical diversity of $>10^{13}$ in humans[2].

Due to the attractive features of antibodies, especially their ability to bind their target specifically with high affinity, their use have spread far beyond their original function in the immune system. They are widely applied in basic research, where they form the basic tool in a variety of methodologies[1]. As such, they have contributed greatly to the scientific knowledge generated in the last couple of decades. Furthermore, they play an important role in diagnostics, often as reagents in ELISA-based tests. Finally, in recent years they have become crucial in disease therapy. New antibody-based therapeutics are approved on a regular basis and the annual revenue is estimated to be at least 20 billion dollars[3]. In summary, antibodies represent molecules of enormous interest and their characterization is of tremendous importance.

An introduction to food allergy

Allergy – an epidemic on the rise

Antibodies are involved when the immune system malfunctions and give rise to e.g. autoimmune diseases and allergy. Allergy is defined as an immunologically mediated hypersensitivity leading to disease. It is often described as an "epidemic" due to its dramatic rise in prevalence across the world. In recent decades the prevalence has continued to rise and currently affects 10-30% of the population[4,5]. The occurrence of allergy correlates with urbanization and hygiene. At first, respiratory allergies (allergic rhinitis (hay fever) and asthma) developed among the wealthiest before spreading to the middle class and finally also affecting the poor[6]. In a similar way, respiratory allergies and eczema are becoming more prevalent in middle income countries, especially in urbanized areas[6]. This has led to a proposal of the 'hygiene hypothesis'[7], today also defined as a 'biodiversity hypothesis'[8]. More recently, food allergies are becoming more prevalent in westernized populations and they have been called the second allergy epidemic[9,10]. An early life exposed to a multitude of antigens, provided by nutrition and infections, seem to be necessary for a proper development of the immune system and a prevention of allergic diseases.

Food allergy – Pathophysiology

Food allergy affects around 5% of young children and 3-4% of the adults in Western countries[11,12]. It is a major problem and the prevalence has been increasing in the last couple of decades[9,10,13–16]. Humans consume many dietary proteins every day, but only a subset of them are allergenic. The allergenic dietary proteins share several characteristics including a relatively small size (< 70 kDa) and resistance to heat and digestion[17]. The mucosal immune system in the intestines regularly encounters non-self antigens and must suppress immune reactivity to food proteins and other harmless antigens, thus developing 'oral tolerance'[11,17]. Antigen-presenting cells, including dendritic cells, and regulatory T cells play a central role in the development of oral tolerance[18,19]. Sometimes the tolerance induction towards a dietary protein fails and the subject becomes 'sensitized'[11,17] (see Figure 1). Allergic sensitization, defined as the production of allergen-specific Immunoglobulin E (IgE)[20], occurs after a B-cell binds an allergen and interact with a CD4+ T cell, which has a specificity for the same allergen[17]. The B cells produce soluble antibodies that are bound by the high-affinity IgE receptor FcεR1 on mast cells and basophils. The elicitation phase of the allergic response occurs when the allergen is re-encountered and

bound by the IgE on mast cells and basophils. This leads to a cross-linking of the FcεR1 receptors, which triggers degranulation of the cells. With the degranulation, immune mediators such as histamines and cytokines are released and these mediators are responsible for a variety of disease symptoms, ranging from itching and vomiting to anaphylaxis[21–24].

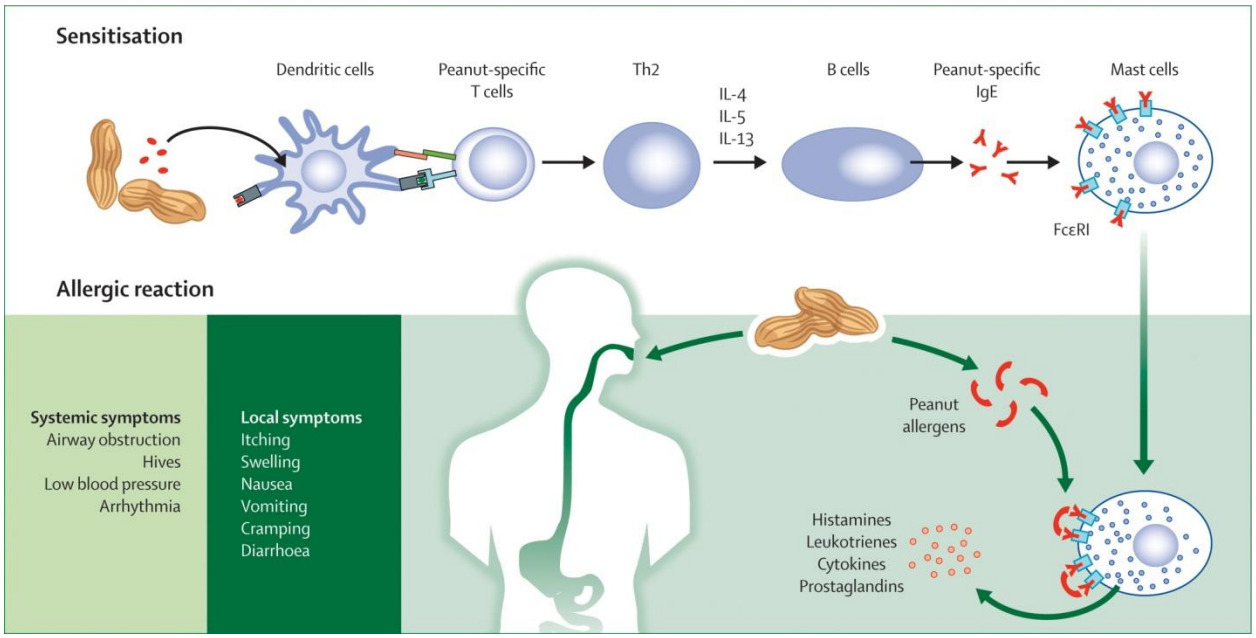


Figure 1. An overview of food allergy pathophysiology. The top section illustrates the principle behind the sensitization phase, whereas the bottom section describes the elucidation phase. The green boxes in the lower left list typical symptoms in food allergy. Peanut allergy is shown as an example, but the general principles are similar for other food allergens. See text for more details. Source: W. Burks, Lancet[25].

Food allergy – Diagnosis & therapy

Food allergy can be diagnosed using several techniques. The golden standard for diagnosis is the double-blind placebo-controlled food challenge (DBPCFC) where the subject is fed small amounts of the suspected food in a controlled environment. However, due to its complexity and the potential risk for the patient it is often replaced by a combination of alternative approaches. Initially, the suspected food may be identified by analyzing the dietary intake history and attempting to associate it temporally with the onset of symptoms[25]. Next, the presence of allergen-specific IgE antibodies can be tested by *in vitro* tests or by skin-prick tests[11]. The findings can finally be supported by food challenges or dietary

interventions[11,26]. It can be complicated to diagnose food allergy, since the detection of specific IgE, e.g. in a positive skin-prick test, does not necessarily correlate with clinical symptoms[11,27,28].

At present, there are no approved treatments for food allergy[29,30]. Instead, the focus is on managing the disease. The primary approach is strict avoidance of the allergen[11,31]. In some cases, this may be supplemented with readily accessible epinephrine to be self-administered in the event of anaphylaxis. In addition, several therapeutic approaches are under investigation with oral immunotherapy as a major strategy[29]. In this approach, the allergenic source is fed in tiny amounts in progressively increasing doses toward a stable level with the aim of inducing desensitization. In immunotherapy, desensitization refers to a protective effect that requires daily consumption of the food allergen in order to maintain protection. The desensitization involves a variety of mechanisms and modifications of the immune response including an up regulation of immunoglobulin G4 (IgG4), which is supposed to act as a "blocking antibody" and compete with IgE[32,33]. So far, studies have demonstrated that immunotherapy induces desensitization but it is not well-established whether tolerance is maintained over a long-term period[34,35].

Peanut allergy

Allergy to peanuts is considered a major food allergy affecting approximately 1-2% of the children in USA[25,36–38]. Within food allergies, it is the leading cause of anaphylaxis and death and it imposes a considerable psychosocial burden on patients and their families[39]. It is generally considered to be a relatively persistent food allergy; however about 20% of the children outgrow their allergy by school age[11,25]. Studies have shown that reduced food diversity in infants' diet is associated with allergic diseases later in childhood[40,41]. A recent comprehensive trial found that dietary introduction of peanuts to infants significantly decreased the risk of peanut allergy[42]. Individuals with peanut allergy often have a history of atopy (atopy is a tendency to be "hypersensitive") and therefore suffer from e.g. asthma, dermatitis and/or rhinitis. The atopic individuals have an increased risk of experiencing life-threatening or fatal allergic reactions to peanuts[43,44].

Molecular peanut allergology

Major peanut allergens

More than half of plant food allergens can be categorized into four structural protein families: the prolamin superfamily, the cupin superfamily, the profilins and Bet v-1-related proteins[45,46]. Almost all of these are either plant defense-related proteins or storage proteins[47]. This is also the case in peanut allergy. An online search at www.allergen.org reveals that, at present, there are 14 recognized allergens (named Ara h 1-15, excluding Ara h 4) and 8 of them can be categorized into the four protein families mentioned above. Specifically, the categories are as follows: the Cupin superfamily for Ara h 1, 3; the Prolamin superfamily for Ara h 2, 6, 7, 9; the Profilin family for Ara h 5; Bet v-1-related proteins for Ara h 8; Oleosin family for Ara h 10, 11, 14, 15; and the Defensin family for Ara h 12, 13.

In North America and Northern Europe Ara h 1, 2 and 3 are considered the dominating peanut allergens[48,49] . In Southern Europe reaction to Ara h 9 is observed at a higher frequency[50]. More recently, Ara h 6 has emerged as an additional important allergen with moderate homology to Ara h 2[51–53]. Ara h 8 appears to be responsible for cross-reactivity to birch pollen allergens[54]. Importantly, the clinical outcome seems to be more associated with the diversity of the allergen recognition rather than recognition of individual proteins[55,56]. The exception is Ara h 2, where recognition is associated with more severe symptoms, especially respiratory symptoms[49].

IgE levels and their correlation with allergy

IgE antibodies and their role in allergy was discovered in the 1960´s[57,58]. IgE is the least abundant antibody isotype with a serum concentration measured in ng/mL[21,59]. As mentioned, the presence of allergen-specific IgE in serum is not necessarily associated with an active allergic disease[11,27,28]. However, several studies have shown that an increased level of specific IgE correlates with a greater risk of clinical symptoms[60,61]. Due to the risks and distress associated with oral food challenges, it is desirable to predict the clinical reactivity based on IgE levels, which has consequently been investigated in numerous studies[60–62]. These studies indicate that it is possible to diagnose allergy based on IgE levels, however, the predictive IgE level vary by food source and between studies. Regarding peanut allergy, a study by Sampson *et al.* reported that there was no correlation between peanut specific IgE level and disease severity[60], however, other contradicting studies did report an association[55,63]. Finally, studies have

examined the ratio of specific IgE to total IgE and they stated that high ratios correlated with the risk of allergy[64,65].

IgG4 – the blocking antibody in allergy

IgG antibodies can be divided into four subclasses of which subclass 4 (IgG4) has attracted a lot of attention due to its association with allergy[32,66,67]. It is the least abundant IgG subclass. Unlike the other IgG subclasses, the concentration of IgG4 varies greatly among healthy people. Whereas the most abundant subclass IgG1 is present in serum in the range of 5-11 mg/ml, IgG4 levels generally range from less than 10µg/ml to 1.4 mg/ml corresponding to 0.8–11.7 % of the total IgG[68]. Besides the variety in concentration, the IgG4 antibodies are also unique in that they undergo 'half-antibody exchange', also known as 'Fab arm exchange'. The result of the exchange is that most IgG4 antibodies in circulation have two different Fab arms making them bi-specific and monovalent for a given antigen[67].

In food allergy, the appearance of specific IgG4 is generally associated with a reduction in symptoms. This is thought to be a consequence of an allergen-blocking effect at the mast cell and basophil level and/or at the level of the antigen-presenting cell thus preventing activation of T cells[66,67,69]. Although the causality remains to be fully unraveled, an increased level of IgG4 has been reported to correlate with the development and maintenance of clinical tolerance; e.g. the desensitization that follows peanut immunotherapy has been associated with increased IgG4 levels[70–72]. This is also the case in other allergies[32]. Although the increase in IgG4 has commonly been associated with a decrease in IgE, there are many studies that did not observe such an association[73]. As a result, it is relevant to determine both the IgG4 and the IgE levels. In a recent trial, where the dietary introduction of peanuts to infants significantly decreased the risk of peanut allergy, the IgG4 levels were also studied[42]. In the group, which consumed peanuts, an earlier and increased level of peanut-specific IgG4 was observed. This up regulation mimics the observations in immunotherapy, reported above. Furthermore, unless peanut-specific IgE levels were very high, elevated IgG4 levels were associated with the absence of an allergic reaction to peanuts[42]. In general, IgG4 is associated with non-activating characteristics in allergy. However, in other pathologies, e.g. pemphigoid diseases and sclerosing diseases, it seems to play a more promoting role[67].

Epitope mapping

Epitopes

The term 'epitope' was coined in 1960 by the Danish immunologist Niels Kaj Jerne[74]. In the context of immunology, an epitope is the part of an antigen that is involved in specific binding to an antibody (B-cell epitope) or T-cell receptor (T-cell epitope). As is the general practice, the terms epitope and B-cell epitope will be used interchangeably unless specified otherwise. The region of the antibody responsible for the specific interaction with an epitope is called a paratope. Epitopes are generally classified into two major groups, conformational or linear epitopes[75]. Conformational epitopes are also known as discontinuous epitopes in which the amino acids involved in antibody binding are located at discontinuous positions in the protein sequence and are brought together by protein folding. In contrast, linear epitopes (also known as continuous or sequential epitopes) are simpler in that they comprise of amino acids that are contiguous in the primary protein sequence.

Although the concept of an epitope is relatively simple, defining or "mapping" the epitope experimentally is a much more challenging task. It is important to recognize the distinction between the structural and functional properties of a given epitope[75–77]. Structural epitopes are defined as all amino acid residues in contact with the paratope without information on the functional importance of these residues. X-ray crystallography can be used to completely define a structural epitope. Conversely, functional epitopes are defined by identifying amino acid residues that are important for binding. Most epitope mapping methods are suited for the study of functional epitopes. Another important notion is the concept of mimotopes. Mimotopes are peptides that differ from the native epitope of an antigen but are still able to interact with the antibody in a similar way[78,79].

Epitopes have been suggested to consist of at least eight amino acid residues, however, energy calculations have suggested that five or six amino acids may be sufficient for antibody binding[80–82]. Kringelum *et al.* analysed antigen:antibody complexes that had been structurally defined by x-ray crystallography and they observed that an average of 15 amino acids were close to the paratope and were therefore potentially involved in the antibody binding or facilitation thereof[83]. They reported that more than half of the epitope comprised of a linear stretch of amino acids. Finally, they described the general B-cell epitope as "a flat, oblong, oval shaped volume consisting of predominantly hydrophobic amino acids in the center flanked by charged residues"[83].

What can we learn from epitope mapping?

Epitope mapping is the process of identifying antibody binding sites. It has a variety of applications. First of all, epitope mapping can provide a fundamental understanding of antibodies and the immune system. Characterization of the epitope-paratope interaction is important in understanding the ways that the immune system reacts during infection, allergic reactions and autoimmune diseases. E.g. in allergy it is important to elucidate what is special about the structure of allergen epitopes that make allergens less tolerable than other proteins[46,84,85]. In addition, discovering epitopes in related proteins may help explain the cross-reactivity often observed in allergy[85].

Regarding diagnostics there have been many attempts at identifying disease biomarkers based on epitope mapping approaches[75]. Furthermore, there have also been an interest in correlating specific epitopes to clinical outcome[11]; see section "Epitope mapping in peanut allergy" for details. However, the greatest clinical interest has probably involved the therapeutic utilization of epitope knowledge. In particular, epitope-based vaccines have received significant attention[75,86–90]. The objective of these efforts is to utilize the structural information to carefully select certain antigenic protein regions and engineer a vaccine based on that. The idea is that the antigen can be more effectively produced and is more homogenous. Furthermore, it is expected that the immune reaction will be more focused and hence more effective, and that the vaccination will have fewer side effects. Epitope mapping can also be used to identify new therapeutic targets[87]. In immunotherapy against allergy, additional therapeutic applications include the use of immune-regulatory T cell epitope peptides[91] as well as molecules tailored to be hypoallergenic[92]. Finally, epitope mapping also plays an important role during the development of therapeutic antibodies. The mapping is used to identify the specific target of an antibody which can aid future development, help in the selection of lead antibody candidates as well as assist in protecting intellectual property rights[89,93].

Epitope mapping techniques

Numerous methods can be applied for epitope mapping. X-ray crystallography is generally considered the golden standard in defining an epitope structure. After successful co-crystallization of the antigen:antibody complex the structure can be determined in great detail based on analysis of the x-ray diffraction patterns[46,86,90,94]. However, x-ray crystallography is a very complex process that is time-consuming and requires a high level of expertise[86]. Consequently, this approach has a low throughput. It

is centered on structural rather than functional information and it is one of the few techniques that actually enable the mapping of conformational epitopes. Another biophysical technique is based on Nuclear Magnetic Resonance (NMR)[46,95]. It provides a more dynamic picture of the antigen:antibody complex in solution. However, like x-ray crystallography it requires extensive expertise and expensive instrumentation[86].

Alternative approaches to epitope mapping are the display techniques, where a large library is screened for binders which can assist in locating the epitope[96]. The display methods can e.g. be based on bacteria[97], yeast[98,99] or bacteriophages[100–102]. The bacteriophage approach, 'phage display', will be covered in detail below. Compared to the biophysical approaches, the display techniques have a relatively high throughput. They are based on functional testing but provide less structural detail. Epitope candidate peptides, often linear candidates, can also be tested for antibody binding on peptide microarrays[103,104]. This technique will also be covered in greater detail in a separate section.

Finally, there are also *in silico* based methods to infer epitope locations[46,105]. These bioinformatic approaches are fast to execute compared to traditional experimental methods. However, the lack of actual experimental data means that the predictions can be relatively inaccurate. As a result, the main use of the *in silico* methods is to narrow down the epitope location which can then be confirmed experimentally[46,86]. B-cell epitopes are more structurally diverse than T-cell epitopes and accordingly they are harder to predict[106].

Phage display

Bacteriophages (or 'phages' in short) are a diverse group of viruses that infect bacteria. Due to their genetic and structural simplicity, as well as the ability to easily grow them on bacterial hosts, phages have been extensively used in basic and applied life science ever since their discovery in the early twentieth century[107,108]. In 1985 George P. Smith reported the first attempt at expressing a foreign peptide on the surface of a phage[109]. This was the start of phage display, which has since grown to become a major display technique applied in a multitude of settings[100,101,110,111].

Highly diverse libraries form the foundation of phage display and they consist of phage particles expressing a great variety of exogenous peptides fused to the phage surface proteins[101,110]. The basic principle of a phage display screening is shown in Figure 2. First, the phage library is incubated with the

target of interest and non-binding phages are washed away. This affinity-selection step is often called 'bio-panning'. Next, the selected phages are eluted e.g. by altering the pH. The eluted phages are then amplified by infecting bacteria. The amplified phages can be used for additional selection rounds. Alternatively, individual phage clones can be picked and the interaction to the target can be confirmed and the clones can be sequenced. The sequencing allows for an identification of the displayed polypeptide in the selected phage.

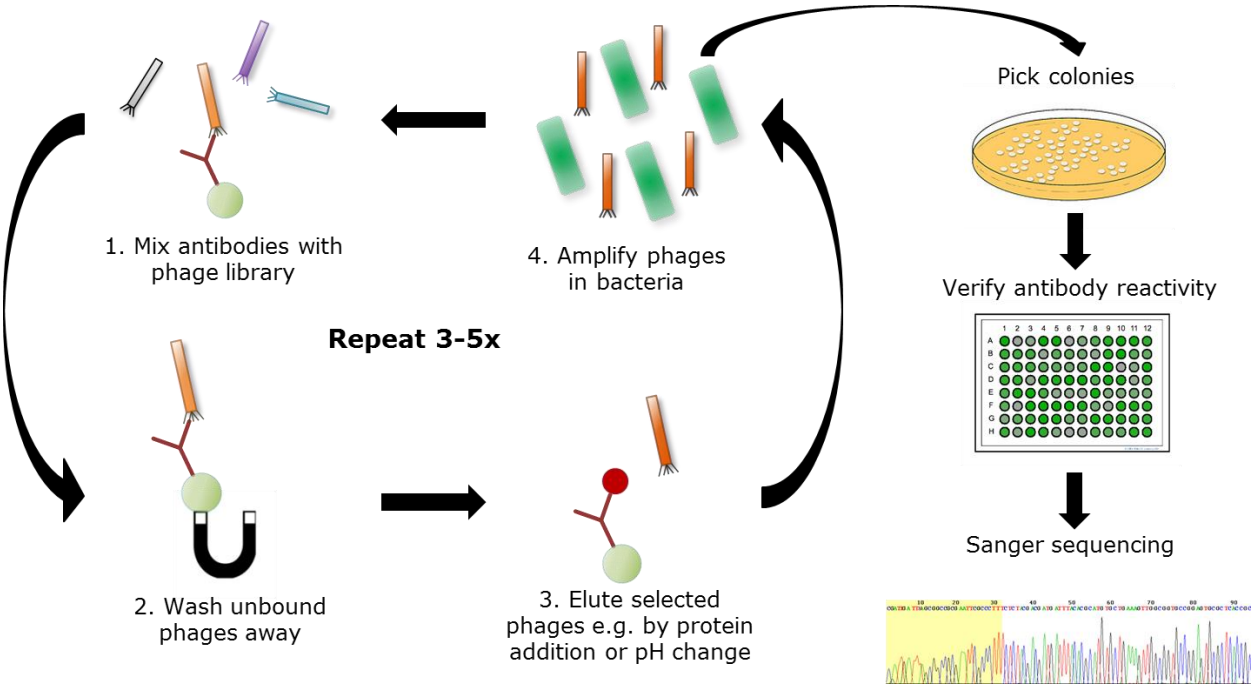


Figure 2. The basic principle of a traditional phage library screening. 1. The phage library is incubated with the immobilized target. As an example, an antibody-target immobilized on magnetic beads is depicted. 2. Unbound phages are washed away. 3. The selected binders are eluted e.g. by a change in pH or competitive addition of a protein; the latter is shown. 4. The eluted phages are amplified by infecting bacteria. The amplified phages can be used for additional selection rounds, starting at 1. Alternatively, amplified individual phage clones can be picked and tested for target reactivity and/or sequenced. The sequencing enables the identification of the nucleotide insert in the selected phage. The figure incorporates elements under the Creative Commons license.

Phage display has been used for various applications[100,101,110,111]. It has been used for general studies of protein-protein interactions[101]. Coupled with combinatorial mutagenesis it provides a rapid method to identify the residues that contribute to binding at a protein-protein interface. Moreover, phage displayed random peptide libraries have been used to identify novel interaction-partners for proteins. In enzyme research, phage display has been used to determine substrate specificity and to develop

enzyme modulators[101,112,113]. Screening of phage display libraries expressing antibody fragments have been widely used to generate specific antibodies[96,101,114,115]. In fact, some of the therapeutic antibodies used today were originally discovered by phage display[116]. Finally, phage display has been used extensively for epitope mapping[75,100–102].

A variety of phage types have been utilized to display a wide array of different polypeptides, however, the filamentous phages are most commonly applied[100]. Of the filamentous phages, the non-lytic M13 is the most frequently used[101]. M13 has a circular genome (6000-8000 bp) encapsulated in a flexible rod coat composed of five different proteins[100,101,117]. The majority of the outer shell consists of several thousand copies of the pVIII protein. One end of the phage particle is capped by 3-5 copies of pVII and pIX and the other end has five copies of pIII and pVI. Foreign peptides have been fused to several of the coat proteins for display, however the fusions to pVIII and especially pIII are the most common[100,101]. The pIII protein is about 405 amino acids long and is involved in phage-host interaction by binding to the bacterial F-pilus during infection. There is a size-limit for the displayed peptides since larger foreign peptides interfere with the coat protein function in viral particle assembly as well as the ability to infect bacteria[101].

For epitope mapping, phage display libraries are either based on random peptides or on fragments from a gene of interest[75]. The reduced diversity of the gene-specific libraries may simplify the process of obtaining a relevant epitope. However, the epitope identification is limited to the gene-specific peptides used as input and a new library has to be constructed for each investigated gene. Mapping of the linear epitopes can be relatively straightforward, especially for gene-specific libraries, as the affinity-selected peptides are similar to linear stretches of the antigen sequence. In contrast, mapping of conformational epitopes provides a greater challenge, and can generally only be accomplished using random libraries. In the latter case, mimotopes mimicking the conformational epitope are obtained in the affinity selection[102,118]. These mimotopes do not necessarily share any sequence similarity with the linear antigen sequence and in order to pinpoint the epitope location the mimotopes are compared with the three-dimensional structure of the antigen. Several bioinformatic tools have been developed to accomplish this, including some that are publicly available[119–121].

Target-unrelated peptides in phage display

An important concern in phage display is the selection of peptides that do not bind the target of interest, collectively known as 'target unrelated peptides' (TUPs)[122–124]. In general, TUPs are suspected to either bind an invariable part of the selection platform or bestow a proliferation advantage to the phage. Invariable parts of the selection platform consist of constant parts of capturing molecules (e.g. antibodies) as well as the solid phase (e.g. plastic wells or beads)[122,123]. As part of the biopanning process, selected phages proliferate in bacteria. During this process, a large phage population propagate in parallel and minor proliferation advantages can be of critical importance. Accordingly, Derda *et al*. have discussed how library diversity is shaped by selection as well as proliferation[124]. Altered proliferation rates of specific phage particles have been suggested to arise by changes in the use of rare nucleotide codons, by interfering with the binding to bacteria pili, by regulatory mutations in phage genes or by obstructing viral packaging or infection[124]. A classic example of a TUP with a propagation advantage is the HAIYPRH peptide which is commonly identified in screens using the popular phage library Ph.D.-7™[123]. In this specific case the propagation advantage is due to a regulatory mutation[125].

There are a number of ways to mitigate the TUP issue. Some approaches aim at decreasing the number of unspecific binders, while others seek to neutralize the proliferation issues[122]. The most prominent way to diminish the number of TUPs is to include experimental steps of negative selection (or deselection)[122,126]. This could e.g. include pre-incubation of the phage library with the solid phase coated with the capturing reagent (e.g. streptavidin). Other attempts at reducing the selection for TUPs include an increase in the coating-density of the target and using Fab fragments instead of whole antibodies in selections that involve antibodies[122]. The purpose of the latter is to decrease unspecific binding to constant parts of the antibody. In order to tackle the TUPs that have propagation advantages, phage selections with no amplification have been attempted with some success[124]. Alternatively, the competition between phages can possibly by decreased by amplifying them on agar plates instead of in liquid culture or by having the amplification occur in small compartments such as microdroplets[124,127,128]. Finally, the issue with TUPs can be partly contended with by comparing the identified phage clones with databases collecting previously isolated phage peptides and known TUP motifs[118,129]. Ultimately, the identified phage peptides have to be validated by independent methods.

Peptide microarrays

The technology of protein microarrays, i.e. immobilized proteins on a surface, have undergone noticeable progress over the last decade and it has been widely applied to study protein-protein interactions[130–132]. One subgroup of the protein microarrays are the peptide microarrays, where peptides (often 10- to 20-mers) are synthesized in parallel, typically on a glass slide or a cellulose membrane. Peptide arrays have a wide range of applications including in drug discovery and enzyme substrate analysis[130,131,133]. As an example of the latter they have been used as a screening tool to identify active kinases[130]. However, a major application of peptide microarrays is in epitope mapping[134] (se e.g. the section below on "Epitope mapping in peanut allergy").

While most standard peptide synthesis equipment can synthesize a few hundred peptides, more recent advances with the SPOT technique have permitted the synthesis of up to 8,000 peptides[135]. However, new advances in miniaturization of the peptide arrays[136] along with developments of an *in situ* photolithic synthesis[137] have recently allowed Buus *et al.*[135], Carmona *et al.*[138], Hansen *et al.*[139], Legutki *et al.*[140] and Forsström *et al.*[134] to synthesize 70,000; 175,000; 200,000; 330,000 and 2,100,000 peptides in parallel, respectively. The Forsström study allowed the spotting of the entire human proteome using 12-mer peptides that had six overlapping residues. It was used to study the specificity and cross-reactivity of monoclonal and polyclonal antibodies. The identified epitopes were 5-7 amino acids long and therefore cross-reactive peptides in off-target proteins were frequently observed. However, their subsequent analysis demonstrated that the off-target peptides were not recognized in a recombinant protein context. The authors speculated that the peptides on microarrays are very flexible and can adapt to the antibody, while the same peptides as part of a protein are more restricted in their conformation[134].

The photolithographic process used for peptide microarray synthesis is shown in Figure 3 along with the subsequent detection of antibodies. The light from a UV-light source is guided through digitally controlled micromirrors to selectively target small spots on the array. The light is used to remove specific photo-sensitive protection groups which are coupled to amino acids on the array. This allows for subsequent binding to additional amino acids with protective groups. Through repeated cycles of lighting and addition of amino acids, the peptides are synthesized on the microarray. After the synthesis, an antibody of interest can be added to the microarray and its location on the array can be detected using secondary antibodies conjugated with fluorophores. Since a given position corresponds to a known peptide, the binding peptide can be determined.

It is possible to identify the general position of an epitope by synthesizing overlapping peptides that cover the target of interest. However, if the epitope is to be characterized at the residue level, a mutagenesis scan has to be included. Mutating the residues of interest to alanine, a so-called 'alanine scan', is a popular way to determine the contribution of specific residues to antibody binding[141]. Alanine is frequently used for the substitutions since it eliminates the side chains without altering the peptide backbone[141]. Studies employing full mutation scans, where each position is mutated to all possible amino acids, have indicated that the alanine scan may sometimes miss semi-important residues[135].

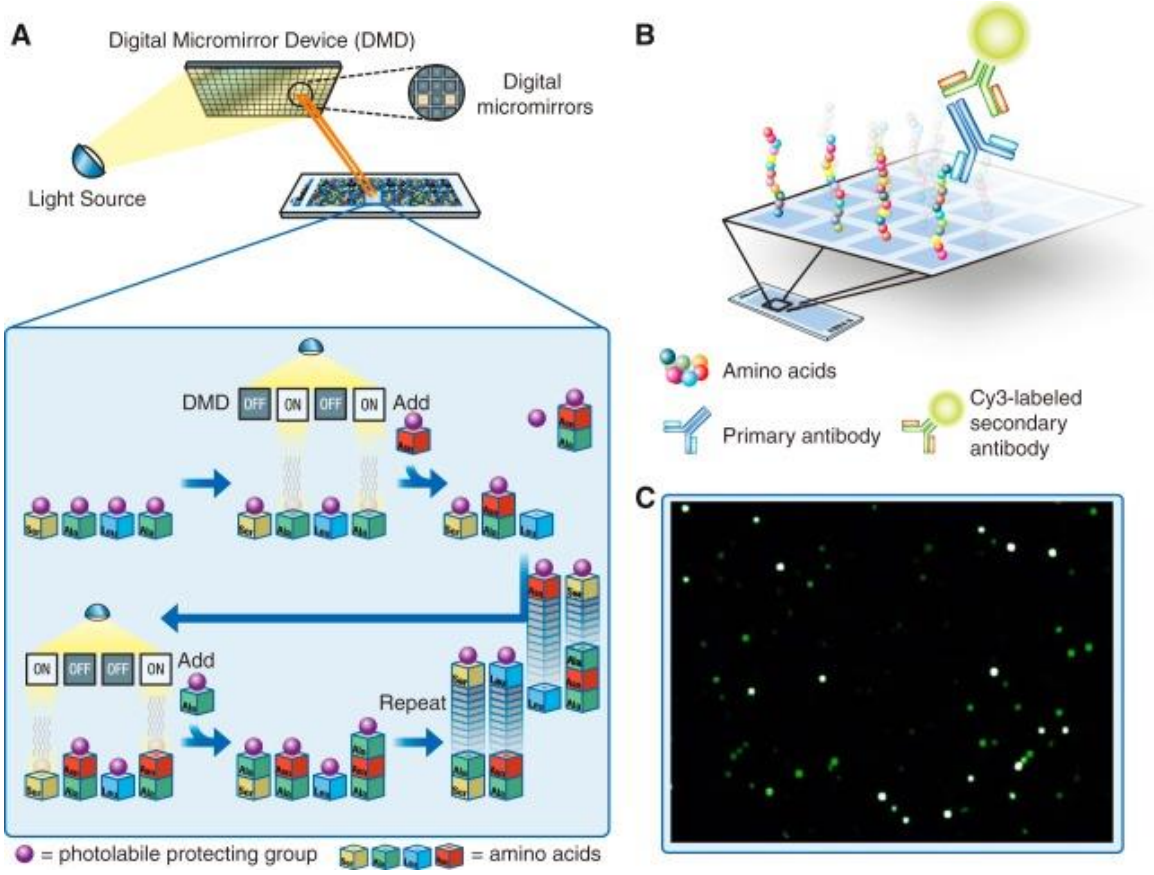


Figure 3. The principle of *in situ* peptide microarray synthesis and antibody binding analysis. **A)** Digitally controlled micromirrors selectively target small areas. The light removes photo-sensitive protection groups from previously bound amino acids. Additional cycles of activation and addition of amino acids synthesizes the peptides of interest. **B)** Incubation of an antibody of interest (primary antibody) and a fluorophore-coupled secondary antibody on the peptide microarray. **C)** An example of an image of a part of the peptide array, where bright spots correspond to peptides bound by antibodies. Source: Forsström *et al.*[134].

Epitope mapping in peanut allergy

Over the past decade, there has been a considerable effort in investigating the role of epitopes in food allergy. The interest has particularly revolved around correlation of epitope patterns with clinical phenotypes, often investigated using peptide microarrays (see Table 1). For peanut allergy, Shreffler *et al.* detected linear IgE epitopes to either Ara h 1, 2, or 3 in 87% of the patients[142]. They observed a noticeable heterogeneity between patients in the number of epitopes as well as their position. In addition, they observed a correlation between the number of epitopes (epitope diversity) and the disease severity. This observation is in line with studies in other allergies. In milk and egg allergy, IgE epitope diversity has been associated with allergy persistence[143–146]. Furthermore, the epitope diversity has been correlated with disease severity in wheat[147] and milk allergy[143]. The correlation with severity was also observed in later peanut studies[148,149]. However, a recent study reported the opposite correlation[150]. Specifically, they observed that patients with severe allergy recognized fewer epitopes. The authors speculate that their normalization, based on specific IgE levels, were responsible for the diverging results.

Based on the correlations between epitope diversity and clinical outcome it has been suggested that specific epitopes can act as biomarkers for various phenotypes[151]. For peanut allergy, Beyer *et al.* initially described that patients with persistent allergy recognized specific IgE epitopes which were not recognized by patients that outgrew their allergy or were peanut sensitized but tolerant[152]. Likewise Lin *et al.* used machine learning to identify 4 peptides (two from Ara h 1 and one from Ara h 2 and Ara h 3) that retrospectively could separate peanut-tolerant from peanut-allergic persons with approximately 90% sensitivity and 95% specificity[149]. Furthermore, Otsu *et al.* clustered the patients based on Ara h 2 and Ara h 6 epitopes and found a tendency towards clustering of clinically related subjects[150]. On the other hand Flinterman *et al.* did not identify any epitopes that could be correlated to disease severity[148].

The studies that also examined IgG4 epitopes have observed that the IgE and IgG4 epitopes overlap[70,148,153]. Specifically, they have reported that 30-50% of the IgE binding peptides were also positive for IgG4 signal. However, these studies have also mentioned that the IgG4 signal is less prominent than the IgE signal[148,149,153]. E.g. Lin *et al.* observed that there was no difference in the IgG4 epitope binding and diversity between peanut-tolerant and peanut-allergic individuals, which was contrary to the observations for IgE[149]. However, they did find a difference in epitope positions between the two patient groups. Along these lines, Flinterman *et al.* reported that epitope diversity of IgE, but not IgG4, correlated with the clinical sensitivity[148]. They also observed that IgE as well as IgG4 epitope

Table 1: Studies of peanut allergy epitopes by peptide microarrays				
Allergens	**Samples**	**Isotypes**	**Main findings in patient sera**	**Reference**
Ara h 1, 2, 3. 20-mer. 3-offset.	77 sera. 15 controls.	IgE	87% had detectable linear epitopes. High heterogeneity in number and patterns of epitope recognition. High epitope diversity correlated with allergy severity.	Shreffler, 2004[142]
Ara h 2. 10/15/20-mer. 2 or 3-offset.	25 sera. 10 controls.	IgE + IgG4	Control and patient sera had different but similar number of IgG4 epitopes. 15- or 20-mer peptides were superior to 10-mers. Overlap in IgE and IgG4 epitopes.	Shreffler, 2005[153]
Ara h 1, 2, 3. 15-mer. 3-offset.	24 children sera. 6 controls.	IgE + IgG4	Positive correlation between IgE epitope diversity and clinical sensitivity. IgG4 signal was less pronounced than IgE. IgE and IgG4 epitopes correlated somewhat and were largely stable over 20 months.	Flinterman, 2008[148]
Ara h 1, 2, 3. 15-mer. 3-offset.	31 children sera. 31 tolerant control sera	IgE + IgG4	Peanut allergic children had higher IgE binding epitope diversity than peanut-tolerant individuals. No significant difference in IgG4 binding was found.	Lin, 2012[149]
Ara h 1, 2, 3. 15-mer. 3-offset.	22 children sera. 6 controls	IgE + IgG4	IgG4 signal increased after OIT and new epitopes emerged. IgE signal generally contracted.	Vickery, 2012[70]
Ara 2, 6. 20-mer. 3-offset.	19 sera w/ severe & 11 sera w/ mild allergy	IgE	Patients with severe allergy recognizes fewer linear epitopes.	Otsu, 2015[150]

binding patterns remained relatively stable over a period of 20 months in children with peanut allergy[148]. Vickery *et al.* looked at the effect of oral immunotherapy (OIT) on epitope patterns[70]. They found that the IgG4 signal generally increased and they observed the emergence of new epitopes. In contrast, IgE signal generally decreased; however, there were individuals in which new specificities were observed. The induction of antigen-specific antibodies has previously been reported in peanut OIT trials[72,154]. The co-localization of IgE and IgG4 epitopes in peanut as well as other allergies has been suggested to be critical for the development of clinical tolerance, based on the idea that specific IgG4 antibodies block the corresponding IgE binding to the allergen[70,155–157].

Sequencing the immune repertoire and managing sequencing errors

The emergence of high-throughput sequencing (HTS) techniques[158] is transforming multiple areas, including immune repertoire characterization[159]. Diversity is a key concept in the immune system; for B-cell receptors the diversification mechanisms can theoretically yield $>10^{13}$ different receptors[2]. However, this exceeds the total number of B lymphocytes in the human body (approximately $1\text{-}2 \times 10^{11}$)[159]. Deep sequencing provides an opportunity to get unprecedented insights into this diverse repertoire. The information gained can be used to develop novel antibody therapeutics and to increase our understanding of the immune system and its behaviour in cancer or after vaccination and infection[159,160].

However, in analyzing diverse repertoires sequencing errors become an important issue. The errors arise in two ways: either during sample preparation (reverse transcription and PCR) or during DNA sequencing[161]. Sequencing errors are an inherent issue in HTS[162–165]. The underlying technology of each sequencing platform influences the type of errors that are observed. The platforms applied in this study were the IonTorrent and the MiSeq platforms. In the IonTorrent platform, the template DNA is amplified in a bead-based emulsion PCR and beads are loaded onto a microchip that detect the change in pH upon the release of hydrogen ions during the sequencing process (so-called ion semiconductor sequencing)[166]. In contrast, for the MiSeq platform, the template is amplified on a solid-phase in flow cells and the sequence is determined based on light emission from fluorescently labelled nucleotides. Whereas the IonTorrent platform is dominated by insertion-deletion (indel) errors, especially issues with determining the exact length of homopolymers, the MiSeq is subjective to substitution errors[165,167,166]. Some error types can be computationally accounted for, at least to some degree[159,164,168]. Substitution error frequencies have been reported to vary from 0.3% to 0.9%[159,166].

Despite the general tendencies, every kind of sequencing error can be observed on any platform. The errors are not only limited to indels or substitutions but also include ambiguous base calls. The sequencing platforms attempt to estimate the error rate for each based called by assigning each base a Phred score[169–171]. The Phred score represents the likelihood that a given base is called incorrectly. For a given base, a Phred score of 20 or 30 represents a likelihood of error of 1% or 0.1%, respectively. The Phred scores also provide a means to ascertain the overall quality of a sequencing run. However, it should be noted that the algorithms used to calculate the Phred score are platform dependent and therefore not directly comparable. E.g. it has been reported that the Illumina platforms have a tendency to overestimate their accuracy whereas the IonTorrent platform, in contrast, underestimates its performance[166,172].

Several approaches may help reduce sequencing errors. E.g. the PacBio platform has been reported to have a much higher accuracy[173,174]. However, this technology is associated with increased expenses as well as an extensive decrease in throughput. An alternative approach to increase the sequencing accuracy is to label each DNA molecule with a nucleotide barcode molecule prior to PCR amplification[175–177]. This makes it possible to count barcodes instead of sequencing reads and enables the elimination of sequencing errors by generating consensus reads having the same barcode. Overall, the barcoding increases the accuracy; however, the sequencing throughput is decreased.

The effort to couple phage display with high-throughput sequencing

Antibody display and gene-specific libraries

Analogously to its use in characterization of the immune repertoire, massive parallel sequencing is well suited to analyse the inherent diversity of phage libraries. Accordingly, recent years have seen considerable efforts in coupling high-throughput sequencing with phage display, often involving the characterization of antibodies or epitopes. Initially, the potential of the approach was demonstrated by studying protein interactions at a grand scale[178–181]. However, with the widespread use of phage libraries that express antibody fragments, analysis of these libraries by HTS has been investigated[182–189] (see Table 2). These studies have demonstrated that considerable insights into the library composition can be gained by HTS analysis. Knowledge of the integrity, size and diversity of a given library is an important aspect in confirming its quality and usability. Furthermore, the HTS data can provide detailed insights into biological parameters such as CDR length distribution and amino acid composition and it provides a

means to ascertain the quality of a library[190]. Deep sequencing also provides a better understanding of the selection process; e.g. how enrichment is correlated with binding affinity and ultimately enables a determination of the optimal number of selection rounds[182,185]. Finally, there is a great interest in enhancing the selection of antibody candidates. Ravn *et al.* pointed out that the approach might make it feasible to by-pass the primary screening and directly focus on in-depth characterization of the most frequent sequences[182]. Acting as proof-of-concept, the coupling with HTS has assisted the identification of relevant antibodies[182,183,189,191]. At a more functional level, Lomakin *et al.* has sought to select cross-reactive antibodies in multiple sclerosis and characterized the selected antibodies in a high-throughput manner by sequencing[187,188]. Additional studies would be required to determine the full potential for this approach in antibody discovery.

Another promising area includes gene-specific libraries. In one study, phages displaying cDNA open reading frames (ORFs) were selected against serum from patients with celiac disease and disease-specific antigens were identified[192]. Another study described the establishment of a display library containing ORFs from secreted microbial proteins for metasecretome studies and characterized the performance by HTS[193]. Scott *et al.* used the IonTorrent platform to support their search for binders using a T7 phage library displaying alpha-1 proteinase inhibitor[194]. Finally, Domina *et al.* recently used a gene-specific library to identify immunodominant regions of a vaccine antigen[195]. However, the resolution of the antibody binding areas was relatively poor, so it was not possible to delineate the epitopes.

Random phage libraries

Just as deep sequencing is highly suited to analyze phage libraries expressing antibodies, described above; it has also been applied in studying the extraordinarily diverse phage libraries that display random peptides (see Table 3). An initial study examined the overlap in peptides obtained by conventional versus HTS-coupled phage display[196]. By comparing large-scale Sanger sequencing data (>3,500 reads) with HTS-derived data they did not observe any difference in GC content. Accordingly, after the translation into peptides, no differences in amino acid distribution was observed. The HTS approach identified 80-95% of the peptides observed in traditional phage display. In contrast, 25-98% of the peptides identified by HTS were not observed with the traditional approach. They concluded that

Table 2: An overview of the studies coupling phage display of antibody fragments with HTS			
Library	**Target**	**Main findings**	**Reference**
scFv w/ CDR3 diversity	5E3, IFNγ	Possible to monitor the selection process of CDR3s. Potent scFv were identified by frequency.	Ravn, 2010[182]
scFv human or murine CDRH3	IFNγ	Antigen biased libraries led to more frequent identification of potent antibodies.	Venet, 2012[183]
Fab from HIV patient	HIV env: gp140	HTS analysis revealed germ line usages, CDR3 length & diversity, somatic mutations and clonally related antibodies	Chen, 2012[184]
scFv w/ 5 CDR3 diversity strategies	5E3, IFNγ	HTS analysis enhanced quality control of complex antibody libraries and facilitated antibody discovery.	Ravn, 2013[185]
scFv Murine CDRH3 from lymphoid organs	CCL5, IFNγ	Antigen biased CDRH3 repertoires increased the performance but reduced the clonal diversity, which may be essential for certain strategies.	Venet, 2013[186]
scFv from Multiple Sclerosis patients	MBP + LMP1	Antibody structures and germlines where highly diverse in cross-reactive clones.	Lomakin, 2013[187]
scFv from Multiple Sclerosis patients	MBP, LMP1, LMP1+MBP, MOG	An overlap of VH but not VL germlines between the MBP and LMP1 sub-libraries were observed.	Lomakin, 2014[188]
scFv w/ CDR3 mutagenesis	ErbB2	Mutant antibodies with increased affinity were identified.	Hu, 2015[189]
scFv artificial design	HER2	Artificially designed library can be used to identify binders.	Chen, 2015[191]

there were no significant differences in the sequencing output between Sanger sequencing and HTS, however, the latter approach provided greater depth.

Matochko *et al.* utilized the HTS capacity to investigate the Ph.D.-7™ phage library diversity after amplification[128]. As expected, certain phages propagated faster than others and a decrease in library diversity was observed. These rapidly propagating phages were suggested to be key TUP candidates. The amplification bias could be significantly reduced if the amplification occurred in emulsion, thus minimizing the competition between phages[128]. This observation is in line with a previous study[197].

With regards to the identification of peptide binders, two studies employed HTS to identify peptides that bound to cell surfaces[198,199]. Hoen *et al.* successfully identified a cell-binding peptide and noted that the peptide was relatively frequent (ranked 14) in the first selection round[198]. Ngubane *et al.* compared the output from HTS with the traditional Sanger approach and observed some overlap, however, the best cell binder was identified using HTS[199]. A study by Rebollo *et al.* investigated the use of HTS in identification of peptide binders by focusing on 5 protein targets with well-described binding motifs[200]. They were able to deduce peptide-binding motifs after a single selection round.

As mentioned, a major application of phage display is in epitope mapping and the potential to integrate HTS with phage display to investigate complex samples, e.g. sera, has been investigated. Ryvkin *et al.* first demonstrated that they could successfully map a known epitope of a monoclonal antibody[201]. They went on to show that biopanning against a pool of HIV positive sera retrieved some phage peptides that suggested HIV specificity. A separate study by Liu *et al.* investigated the similarity between phage peptides selected against serum and all known proteins (using BLAST) and attempted to identify potential IgG binding targets[202]. For one out of two antigens used to immunize mice, alignment of the

peptides obtained after selection against serum suggested the antigen as one of the top candidates. An ELISA test of the suggested epitope area confirmed the antibody reactivity. However, for the second investigated target, the approach did not work. The authors suggested that any epitopes for this antigen were conformational[202]. While the main focus has been on integrating HTS with phage display, recently there have been attempts at coupling it with other display techniques, including ribosome display[203], yeast display[204,205] and mRNA display[206,207].

Table 3: An overview of the studies that have coupled phage display of random peptides with HTS			
Library	**Samples**	**Main findings**	**Reference**
RGD-4C and cyclic 7-mer peptides	4 tissue samples	HTS yields comparable results to traditional phage display.	Dias-Neto, 2009[196]
7-mer peptides	Pool of HIV sera	Monoclonal antibody epitope was successfully mapped. Peptides selected against a HIV sera pool could be aligned to an HIV protein.	Ryvkin, 2012[201]
7-mer peptides	Osteoblast cells	Identified a cell binding peptide which was relatively frequent (ranked 14) in the first selection round.	'Hoen, 2012[198]
12-mer peptides	N/A	Amplification in bacteria leads to a decrease in library diversity.	Matochko, 2012[197]
7-mer peptides	Sera from immunized mice & from 1 melanoma patient	BLAST of selected peptides returned the immunization antigen among top hits. Peptides clustered around a specific position in the antigen. A putative antigen was identified in the patient serum.	Liu, 2013[202]
Cyclic 7-mer peptides	Mycobacterium tuberculosis	Observed overlap in output from traditional and HTS-enhanced phage display, however, the best cell binder was identified using HTS.	Ngubane, 2013[199]
7-mer peptides	N/A	Identified 770 phages with propagation advantages. This bias could be minimized if the phages were amplified in emulsion.	Matochko, 2014[128]
Cyclical 7-10-mer peptides	5 protein targets	Target-binding peptide motifs could be identified after a single selection round.	Rebollo, 2014[200]

References

1. Kenneth Murphy, Paul Travers, M. W. *Janeway's Immunobiology*. (Garland Science, 2008).

2. Schroeder, H. W. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* **30,** 119–35 (2006).

3. Liu, J. K. H. The history of monoclonal antibody development – Progress, remaining challenges and future innovations. *Ann. Med. Surg.* **3,** 113–116 (2014).

4. Ring, J. *et al.* Davos declaration: allergy as a global problem. *Allergy* **67,** 141–3 (2012).

5. Pawankar, R., Canonica, G., Holgate, S. & Lockey, R. WAO, white book on allergy. *Pediatriya* **52,** 55–58 (2012).

6. Matricardi, P. M. 99th Dahlem conference on infection, inflammation and chronic inflammatory disorders: controversial aspects of the 'hygiene hypothesis'. *Clin. Exp. Immunol.* **160,** 98–105 (2010).

7. Yazdanbakhsh, M., Kremsner, P. G. & van Ree, R. Allergy, parasites, and the hygiene hypothesis. *Science* **296,** 490–4 (2002).

8. Haahtela, T. *et al.* The biodiversity hypothesis and allergic disease: world allergy organization position statement. *World Allergy Organ. J.* **6,** 3 (2013).

9. Prescott, S. & Allen, K. J. Food allergy: Riding the second wave of the allergy epidemic. *Pediatr. Allergy Immunol.* **22,** 155–160 (2011).

10. Burney, P. G. J. *et al.* The prevalence and distribution of food sensitization in European adults. *Allergy Eur. J. Allergy Clin. Immunol.* **69,** 365–371 (2014).

11. Sicherer, S. H. & Sampson, H. A. Food allergy. *J. Allergy Clin. Immunol.* **125,** S116–S125 (2010).

12. Wang, J. & Sampson, H. A. Food allergy: recent advances in pathophysiology and treatment. *Allergy. Asthma Immunol. Res.* **1,** 19–29 (2009).

13. Sicherer, S. H., Muñoz-Furlong, A., Godbold, J. H. & Sampson, H. A. US prevalence of self-reported peanut, tree nut, and sesame allergy: 11-year follow-up. *J. Allergy Clin. Immunol.* **125,** 1322–1326 (2010).

14. Hourihane, J. O. Peanut allergy. *Pediatr. Clin. North Am.* **58,** 445–58 (2011).

15. Branum, A. M. & Lukacs, S. L. Food allergy among children in the United States. *Pediatrics* **124,** 1549–1555 (2009).

16. Poulos, L. M., Waters, A. M., Correll, P. K., Loblay, R. H. & Marks, G. B. Trends in hospitalizations for anaphylaxis, angioedema, and urticaria in Australia, 1993-1994 to 2004-2005. *J. Allergy Clin. Immunol.* **120,** 878–884 (2007).

17. Vickery, B. P., Chin, S. & Burks, A. W. Pathophysiology of food allergy. *Pediatr. Clin. North Am.* **58,** 363–76, ix–x (2011).

18. Chehade, M. & Mayer, L. Oral tolerance and its relation to food hypersensitivities. *J. Allergy Clin. Immunol.* **115,** 3–12; quiz 13 (2005).

19. Strobel, S. & Mowat, A. M. Oral tolerance and allergic responses to food proteins. *Curr. Opin. Allergy Clin. Immunol.* **6,** 207–213 (2006).

20. Geha, R. S., Jabara, H. H. & Brodeur, S. R. The regulation of immunoglobulin E class-switch recombination. *Nat. Rev. Immunol.* **3,** 721–32 (2003).

21. Gould, H. J. *et al.* The biology of IGE and the basis of allergic disease. *Annu. Rev. Immunol.* **21,** 579–628 (2003).

22. Stone, K. D., Prussin, C. & Metcalfe, D. D. IgE, mast cells, basophils, and eosinophils. *J. Allergy Clin. Immunol.* **125,** S73–S80 (2010).

23. Winter, W. E., Hardt, N. S. & Fuhrman, S. Immunoglobulin E: importance in parasitic infections and hypersensitivity responses. *Arch. Pathol. Lab. Med.* **124,** 1382–5 (2000).

24. Van Wijk, F. & Knippels, L. Initiating mechanisms of food allergy: Oral tolerance versus allergic sensitization. *Biomed. Pharmacother.* **61,** 8–20 (2007).

25. Burks, A. W. Peanut allergy. *Lancet* **371,** 1538–1546 (2008).

26. Bruijnzeel-Koomen, C. *et al.* Adverse reactions to food. European Academy of Allergology and Clinical Immunology Subcommittee. *Allergy* **50,** 623–35 (1995).

27. Asero, R. *et al.* IgE-mediated food allergy diagnosis: Current status and new perspectives. *Mol. Nutr. Food Res.* **51,** 135–147 (2007).

28. Hamilton, R. G., MacGlashan, D. W. & Saini, S. S. IgE antibody-specific activity in human allergic disease. *Immunol. Res.* **47,** 273–284 (2010).

29. Berin, M. C. & Sicherer, S. Food allergy: Mechanisms and therapeutics. *Curr. Opin. Immunol.* **23,** 794–800 (2011).

30. Nowak-Wgrzyn, A. & Sampson, H. A. Future therapies for food allergies. *J. Allergy Clin. Immunol.* **127,** 558–573 (2011).

31. Cianferoni, A. & Spergel, J. M. Food allergy: review, classification and diagnosis. *Allergol Int* **58,** 457–466 (2009).

32. Matsuoka, T., Shamji, M. H. & Durham, S. R. Allergen immunotherapy and tolerance. *Allergol. Int.* **62,** 403–13 (2013).

33. Jutel, M. *et al.* Mechanisms of allergen-specific immunotherapy and novel ways for vaccine development. *Allergol. Int.* **62,** 425–33 (2013).

34. Sheikh, S. Z. & Burks, a W. Recent advances in the diagnosis and therapy of peanut allergy. *Expert Rev. Clin. Immunol.* **9,** 551–60 (2013).

35. Wang, J. & Sampson, H. A. Oral and sublingual immunotherapy for food allergy. *Asian Pacific J. Allergy Immunol.* **31,** 198–209 (2013).

36. Sampson, H. A. 9. Food allergy. *J. Allergy Clin. Immunol.* **111,** S540–7 (2003).

37. Sampson, H. A. Update on food allergy. *J. Allergy Clin. Immunol.* **113,** 805–819 (2004).

38. Liu, A. H. *et al.* National prevalence and risk factors for food allergy and relationship to asthma: results from the National Health and Nutrition Examination Survey 2005-2006. *J. Allergy Clin. Immunol.* **126,** 798–806.e13 (2010).

39. Cummings, A. J., Knibb, R. C., King, R. M. & Lucas, J. S. The psychosocial impact of food allergy and food hypersensitivity in children, adolescents and their families: a review. *Allergy* **65,** 933–945 (2010).

40. Nwaru, B. I. *et al.* Food diversity in infancy and the risk of childhood asthma and allergies. *J. Allergy Clin. Immunol.* **133,** 1084–1091 (2014).

41. Wang, M. *et al.* Reduced diversity in the early fecal microbiota of infants with atopic eczema. *J. Allergy Clin. Immunol.* **121,** 129–134 (2008).

42. Du Toit, G. *et al.* Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy. *N. Engl. J. Med.* **372,** 150223141105002 (2015).

43. Bock, S. A., Muoz-Furlong, A. & Sampson, H. A. Fatalities due to anaphylactic reactions to foods. *J. Allergy Clin. Immunol.* **107,** 191–193 (2001).

44. Sampson, H. A., Mendelson, L. & Rosen, J. P. Fatal and near-fatal anaphylactic reactions to food in children and adolescents. *N. Engl. J. Med.* **327,** 380–384 (1992).

45. Radauer, C. & Breiteneder, H. Evolutionary biology of plant food allergens. *J. Allergy Clin. Immunol.* **120,** 518–525 (2007).

46. Dall'antonia, F., Pavkov-Keller, T., Zangger, K. & Keller, W. Structure of allergens and structure based epitope predictions. *Methods* **66,** 3–21 (2014).

47. Breiteneder, H. & Radauer, C. A classification of plant food allergens. *J. Allergy Clin. Immunol.* **113,** 821–830; quiz 831 (2004).

48. Nicolaou, N. *et al.* Allergy or tolerance in children sensitized to peanut: Prevalence and differentiation using component-resolved diagnostics. *J. Allergy Clin. Immunol.* **125,** 191–197.e13 (2010).

49. Asarnoj, A. *et al.* IgE to peanut allergen components: Relation to peanut symptoms and pollen sensitization in 8-year-olds. *Allergy Eur. J. Allergy Clin. Immunol.* **65,** 1189–1195 (2010).

50. Krause, S. *et al.* Lipid transfer protein (Ara h 9) as a new peanut allergen relevant for a Mediterranean allergic population. *J. Allergy Clin. Immunol.* **124,** 771–8.e5 (2009).

51. Flinterman, A. E. *et al.* Children with peanut allergy recognize predominantly Ara h2 and Ara h6, which remains stable over time. *Clin. Exp. Allergy* **37,** 1221–1228 (2007).

52. Koid, A. E. *et al.* Ara h 6 complements Ara h 2 as an important marker for IgE reactivity to peanut. *J. Agric. Food Chem.* **62,** 206–213 (2014).

53. Porterfield, H. S. *et al.* Effector activity of peanut allergens: a critical role for Ara h 2, Ara h 6, and their variants. *Clin. Exp. Allergy* **39,** 1099–108 (2009).

54. Mittag, D. *et al.* Ara h 8, a Bet v 1-homologous allergen from peanut, is a major allergen in patients with combined birch pollen and peanut allergy. *J. Allergy Clin. Immunol.* **114,** 1410–7 (2004).

55. Lewis, S. A., Grimshaw, K. E. C., Warner, J. O. & Hourihane, J. O. B. The promiscuity of immunoglobulin E binding to peanut allergens, as determined by Western blotting, correlates with the severity of clinical symptoms. *Clin. Exp. Allergy* **35,** 767–773 (2005).

56. Astier, C. *et al.* Predictive value of skin prick tests using recombinant allergens for diagnosis of peanut allergy. *J. Allergy Clin. Immunol.* **118,** 250–256 (2006).

57. Ishizaka, K., Ishizaka, T. & Hornbrook, M. M. Physico-chemical properties of human reaginic antibody. IV. Presence of a unique immunoglobulin as a carrier of reaginic activity. *J. Immunol.* **97,** 75–85 (1966).

58. Ishizaka, K. & Ishizaka, T. Biological function of gamma E antibodies and mechanisms of reaginic hypersensitivity. *Clin. Exp. Immunol.* **6,** 25–42 (1970).

59. Gadermaier, E., Levin, M., Flicker, S. & Ohlin, M. The human IgE repertoire. *Int. Arch. Allergy Immunol.* **163,** 77–91 (2014).

60. Sampson, H. A. & Ho, D. G. Relationship between food-specific IgE concentrations and the risk of positive food challenges in children and adolescents. *J. Allergy Clin. Immunol.* **100,** 444–51 (1997).

61. Celik-Bilgili, S. *et al.* The predictive value of specific immunoglobulin E levels in serum for the outcome of oral food challenges. *Clin. Exp. Allergy* **35,** 268–73 (2005).

62. Sampson, H. A. Utility of food-specific IgE concentrations in predicting symptomatic food allergy. *J. Allergy Clin. Immunol.* **107,** 891–6 (2001).

63. Hourihane, J. O. *et al.* Does severity of low-dose, double-blind, placebo-controlled food challenges reflect severity of allergic reactions to peanut in the community? *Clin. Exp. Allergy* **35,** 1227–33 (2005).

64. Mehl, A. *et al.* Utility of the ratio of food-specific IgE/total IgE in predicting symptomatic food allergy in children. *Allergy* **60,** 1034–9 (2005).

65. El-Khouly, F., Lewis, S. A., Pons, L., Burks, A. W. & Hourihane, J. O. IgG and IgE avidity characteristics of peanut allergic individuals. *Pediatr. Allergy Immunol.* **18,** 607–13 (2007).

66. Akdis, C. A. & Akdis, M. Mechanisms of allergen-specific immunotherapy and immune tolerance to allergens. *World Allergy Organ. J.* **8,** 17 (2015).

67. Aalberse, R. C., Stapel, S. O., Schuurman, J. & Rispens, T. Immunoglobulin G4: an odd antibody. *Clin. Exp. Allergy* **39,** 469–77 (2009).

68. Nirula, A., Glaser, S. M., Kalled, S. L., Taylor, F. R. & Taylora, F. R. What is IgG4? A review of the biology of a unique immunoglobulin subtype. *Curr. Opin. Rheumatol.* **23,** 119–24 (2011).

69. Santos, A. F. *et al.* IgG4 inhibits peanut-induced basophil and mast cell activation in peanut-tolerant children sensitized to peanut major allergens. *J. Allergy Clin. Immunol.* **135,** 1249–56 (2015).

70. Vickery, B. P. *et al.* Peanut oral immunotherapy modifies IgE and IgG4 responses to major peanut allergens. *J. Allergy Clin. Immunol.* **131,** 128–34.e1–3 (2013).

71. Kim, E. H. *et al.* Sublingual immunotherapy for peanut allergy: Clinical and immunologic evidence of desensitization. *J. Allergy Clin. Immunol.* **127,** 640–646.e1 (2011).

72. Jones, S. M. *et al.* Clinical efficacy and immune regulation with peanut oral immunotherapy. *J. Allergy Clin. Immunol.* **124,** 292–300, 300.e1–97 (2009).

73. Bøgh, K. L. *et al.* IgE versus IgG4 epitopes of the peanut allergen Ara h 1 in patients with severe allergy. *Mol. Immunol.* **58,** 169–176 (2013).

74. Jerne, N. K. Immunological speculations. *Annu. Rev. Microbiol.* **14,** 341–58 (1960).

75. Wang, L.-F. & Yu, M. Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Curr. Drug Targets* **5,** 1–15 (2004).

76. Van Regenmortel, M. H. Structural and functional approaches to the study of protein antigenicity. *Immunol. Today* **10,** 266–72 (1989).

77. Aalberse, R. C. & Crameri, R. IgE-binding epitopes: a reappraisal. *Allergy* **66,** 1261–74 (2011).

78. Geysen, H. M., Rodda, S. J. & Mason, T. J. The delineation of peptides able to mimic assembled epitopes. *Ciba Found. Symp.* **119,** 130–49 (1986).

79. Geysen, H. M., Rodda, S. J. & Mason, T. J. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Mol. Immunol.* **23,** 709–15 (1986).

80. Van Regenmortel MHV. Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. *Methods* **9,** 465–72 (1996).

81. Bannon, G. A. & Ogawa, T. Evaluation of available IgE-binding epitope data and its utility in bioinformatics. *Mol. Nutr. Food Res.* **50,** 638–44 (2006).

82. Laver, W. G., Air, G. M., Webster, R. G. & Smith-Gill, S. J. Epitopes on protein antigens: misconceptions and realities. *Cell* **61,** 553–6 (1990).

83. Kringelum, J. V., Nielsen, M., Padkjær, S. B. & Lund, O. Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol. Immunol.* **53,** 24–34 (2013).

84. Taylor, S. L. & Hefle, S. L. Will genetically modified foods be allergenic? *J. Allergy Clin. Immunol.* **107,** 765–771 (2001).

85. Mueller, G. A., Maleki, S. J. & Pedersen, L. C. The molecular basis of peanut allergy. *Curr. Allergy Asthma Rep.* **14,** 429 (2014).

86. Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Freund, N. T. & Weiss, Y. Epitope mapping: The first step in developing epitope-based vaccines. *BioDrugs* **21,** 145–156 (2007).

87. De Groot, A. S. Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov. Today* **11,** 203–9 (2006).

88. Dormitzer, P. R., Ulmer, J. B. & Rappuoli, R. Structure-based antigen design: a strategy for next generation vaccines. *Trends Biotechnol.* **26,** 659–67 (2008).

89. Sharon, J., Rynkiewicz, M. J., Lu, Z. & Yang, C.-Y. Discovery of protective B-cell epitopes for development of antimicrobial vaccines and antibody therapeutics. *Immunology* **142,** 1–23 (2014).

90. Malito, E., Carfi, A. & Bottomley, M. J. Protein Crystallography in Vaccine Research and Development. *Int. J. Mol. Sci.* **16,** 13106–13140 (2015).

91. Prickett, S. R., Rolland, J. M. & O'Hehir, R. E. Immunoregulatory T-cell epitope peptides: the new frontier in allergy therapy. *Clin. Exp. Allergy* **45,** 1015–26 (2015).

92. Valenta, R., Campana, R., Marth, K. & van Hage, M. Allergen-specific immunotherapy: from therapeutic vaccines to prophylactic approaches. *J. Intern. Med.* **272,** 144–57 (2012).

93. Beck, A., Wurch, T., Bailly, C. & Corvaia, N. Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.* **10,** 345–52 (2010).

94. Pomés, A. Relevant B cell epitopes in allergic disease. *Int. Arch. Allergy Immunol.* **152,** 1–11 (2010).

95. Takeuchi, K. & Wagner, G. NMR studies of protein interactions. *Curr. Opin. Struct. Biol.* **16,** 109–17 (2006).

96. Sergeeva, A., Kolonin, M. G., Molldrem, J. J., Pasqualini, R. & Arap, W. Display technologies: Application for the discovery of drug and gene delivery agents. *Adv. Drug Deliv. Rev.* **58,** 1622–1654 (2006).

97. Daugherty, P. S. Protein engineering with bacterial display. *Curr. Opin. Struct. Biol.* **17,** 474–80 (2007).

98. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17,** 467–73 (2007).

99. Pepper, L. R., Cho, Y. K., Boder, E. T. & Shusta, E. V. A decade of yeast surface display technology: where are we now? *Comb. Chem. High Throughput Screen.* **11,** 127–34 (2008).

100. Smith, G. P. & Petrenko, V. A. Phage Display. *Chem. Rev.* **97,** 391–410 (1997).

101. Pande, J., Szewczyk, M. & Grover, A. Phage display: concept, innovations, applications and future. *Biotechnol. Adv.* **28,** 849–58 (2010).

102. Rowley, M. J., O'Connor, K. & Wijeyewickrema, L. Phage display for epitope determination: A paradigm for identifying receptor-ligand interactions. *Biotechnol. Annu. Rev.* **10,** 151–188 (2004).

103. Uttamchandani, M. & Yao, S. Q. Peptide microarrays: next generation biochips for detection, diagnostics and high-throughput screening. *Curr. Pharm. Des.* **14,** 2428–38 (2008).

104. Stoevesandt, O., Taussig, M. J. & He, M. Protein microarrays: high-throughput tools for proteomics. *Expert Rev. Proteomics* **6,** 145–57 (2009).

105. Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O. & Rosales-Mendoza, S. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J. Biomed. Inform.* **53,** 405–14 (2015).

106. Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Comput. Biol.* **8,** e1002829 (2012).

107. Sulakvelidze, A., Alavidze, Z. & Morris, J. G. Bacteriophage therapy. *Antimicrob. Agents Chemother.* **45,** 649–59 (2001).

108. Pennazio, S. The origin of phage virology. *Riv. Biol.* **99,** 103–29 (2006).

109. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228,** 1315–7 (1985).

110. Bratkovič, T. Progress in phage display: Evolution of the technique and its applications. *Cell. Mol. Life Sci.* **67,** 749–767 (2010).

111. Sidhu, S. S. & Koide, S. Phage display for engineering and analyzing protein interaction interfaces. *Curr. Opin. Struct. Biol.* **17,** 481–487 (2007).

112. Kay, B. K. & Hamilton, P. T. Identification of enzyme inhibitors from phage-displayed combinatorial peptide libraries. *Comb. Chem. High Throughput Screen.* **4,** 535–43 (2001).

113. Deperthes, D. Phage display substrate: a blind method for determining protease specificity. *Biol. Chem.* **383,** 1107–12 (2002).

114. Hoogenboom, H. R. Overview of antibody phage-display technology and its applications. *Methods Mol. Biol.* **178,** 1–37 (2002).

115. Conrad, U. & Scheller, J. Considerations on antibody-phage display methodology. *Comb. Chem. High Throughput Screen.* **8,** 117–26 (2005).

116. Schirrmann, T., Meyer, T., Schütte, M., Frenzel, A. & Hust, M. Phage display for the generation of antibodies for proteome research, diagnostics and therapy. *Molecules* **16,** 412–26 (2011).

117. Kehoe, J. W. & Kay, B. K. Filamentous phage display in the new millennium. *Chem. Rev.* **105,** 4056–72 (2005).

118. Huang, J. *et al.* MimoDB 2.0: A mimotope database and beyond. *Nucleic Acids Res.* **40,** D271–7 (2012).

119. Mayrose, I. *et al.* Epitope mapping using combinatorial phage-display libraries: A graph-based algorithm. *Nucleic Acids Res.* **35,** 69–78 (2007).

120. Bublil, E. M. *et al.* Stepwise prediction of conformational discontinuous B-cell epitopes using the mapitope algorithm. *Proteins Struct. Funct. Genet.* **68,** 294–304 (2007).

121. Huang, J., Ru, B. & Dai, P. Bioinformatics resources and tools for phage display. *Molecules* **16,** 694–709 (2011).

122. Menendez, A. & Scott, J. K. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. *Anal. Biochem.* **336,** 145–57 (2005).

123. Vodnik, M., Zager, U., Strukelj, B. & Lunder, M. Phage display: Selecting straws instead of a needle from a haystack. *Molecules* **16,** 790–817 (2011).

124. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16,** 1776–1803 (2011).

125. Brammer, L. A. *et al.* A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site. *Anal. Biochem.* **373,** 88–98 (2008).

126. Newton-Northup, J. R. Contending With Target Unrelated Peptides from Phage Display. *J. Mol. Imaging Dyn.* **02,** 2–4 (2013).

127. Derda, R., Tang, S. K. Y. & Whitesides, G. M. Uniform amplification of phage with different growth characteristics in individual compartments consisting of monodisperse droplets. *Angew. Chemie - Int. Ed.* **49,** 5301–5304 (2010).

128. Matochko, W. L., Cory Li, S., Tang, S. K. Y. & Derda, R. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res.* **42,** 1784–98 (2014).

129. Huang, J., Ru, B., Li, S., Lin, H. & Guo, F. B. SAROTUP: Scanner and reporter of target-unrelated peptides. *J. Biomed. Biotechnol.* **2010,** 101932 (2010).

130. Wolf-Yadlin, A., Sevecka, M. & MacBeath, G. Dissecting protein function and signaling using protein microarrays. *Curr. Opin. Chem. Biol.* **13,** 398–405 (2009).

131. Berrade, L., Garcia, A. E. & Camarero, J. A. Protein microarrays: novel developments and applications. *Pharm. Res.* **28,** 1480–99 (2011).

132. Weinrich, D., Jonkheijm, P., Niemeyer, C. M. & Waldmann, H. Applications of protein biochips in biomedical and biotechnological research. *Angew. Chem. Int. Ed. Engl.* **48,** 7744–51 (2009).

133. Breitling, F., Nesterov, A., Stadler, V., Felgenhauer, T. & Bischoff, F. R. High-density peptide arrays. *Mol. Biosyst.* **5,** 224–34 (2009).

134. Forsström, B. *et al.* Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays. *Mol. Cell. Proteomics* **13,** 1585–97 (2014).

135. Buus, S. *et al.* High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Mol. Cell. Proteomics* **11,** 1790–800 (2012).

136. Pellois, J. P. *et al.* Individually addressable parallel peptide synthesis on microchips. *Nat. Biotechnol.* **20,** 922–6 (2002).

137. Fodor, S. P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251,** 767–73 (1991).

138. Carmona, S. J. *et al.* Towards high-throughput immunomics for infectious diseases: use of next-generation peptide microarrays for rapid discovery and mapping of antigenic determinants. *Mol. Cell. Proteomics* **14,** 1871–84 (2015).

139. Hansen, L. B., Buus, S. & Schafer-Nielsen, C. Identification and Mapping of Linear Antibody Epitopes in Human Serum Albumin Using High-Density Peptide Arrays. *PLoS One* **8,** e68902 (2013).

140. Legutki, J. B. *et al.* Scalable high-density peptide arrays for comprehensive health monitoring. *Nat. Commun.* **5,** 4785 (2014).

141. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244,** 1081–5 (1989).

142. Shreffler, W. G., Beyer, K., Chu, T. H. T., Burks, A. W. & Sampson, H. A. Microarray immunoassay: Association of clinical history, in vitro IgE function, and heterogeneity of allergenic peanut epitopes. *J. Allergy Clin. Immunol.* **113,** 776–782 (2004).

143. Wang, J. *et al.* Correlation of IgE/IgG4 milk epitopes and affinity of milk-specific IgE antibodies with different phenotypes of clinical milk allergy. *J. Allergy Clin. Immunol.* **125,** 695–702, 702.e1–702.e6 (2010).

144. Järvinen, K.-M. *et al.* B-cell epitopes as a screening instrument for persistent cow's milk allergy. *J. Allergy Clin. Immunol.* **110,** 293–7 (2002).

145. Chatchatee, P. *et al.* Identification of IgE and IgG binding epitopes on beta- and kappa-casein in cow's milk allergic patients. *Clin. Exp. Allergy* **31,** 1256–62 (2001).

146. Järvinen, K. M., Chatchatee, P., Bardina, L., Beyer, K. & Sampson, H. A. IgE and IgG binding epitopes on alpha-lactalbumin and beta-lactoglobulin in cow's milk allergy. *Int. Arch. Allergy Immunol.* **126,** 111–8 (2001).

147. Battais, F. *et al.* Identification of IgE-binding epitopes on gliadins for patients with food allergy to wheat. *Allergy* **60,** 815–21 (2005).

148. Flinterman, A. E. *et al.* Peanut epitopes for IgE and IgG4 in peanut-sensitized children in relation to severity of peanut allergy. *J. Allergy Clin. Immunol.* **121,** 737–743.e10 (2008).

149. Lin, J. *et al.* A bioinformatics approach to identify patients with symptomatic peanut allergy using peptide microarray immunoassay. *J. Allergy Clin. Immunol.* **129,** 1321–1328.e5 (2012).

150. Otsu, K., Guo, R. & Dreskin, S. C. Epitope analysis of Ara h 2 and Ara h 6: characteristic patterns of IgE-binding fingerprints among individuals with similar clinical histories. *Clin. Exp. Allergy* **45,** 471–84 (2015).

151. Lin, J. & Sampson, H. A. The role of immunoglobulin E-binding epitopes in the characterization of food allergy. *Curr. Opin. Allergy Clin. Immunol.* **9,** 357–363 (2009).

152. Beyer, K. *et al.* Measurement of peptide-specific IgE as an additional tool in identifying patients with clinical reactivity to peanuts. *J. Allergy Clin. Immunol.* **112,** 202–7 (2003).

153. Shreffler, W. G., Lencer, D. a., Bardina, L. & Sampson, H. a. IgE and IgG4 epitope mapping by microarray immunoassay reveals the diversity of immune response to the peanut allergen, Ara h 2. *J. Allergy Clin. Immunol.* **116,** 893–899 (2005).

154. Varshney, P. *et al.* A randomized controlled study of peanut oral immunotherapy: clinical desensitization and modulation of the allergic response. *J. Allergy Clin. Immunol.* **127,** 654–60 (2011).

155. James, L. K. *et al.* Allergen specificity of IgG(4)-expressing B cells in patients with grass pollen allergy undergoing immunotherapy. *J. Allergy Clin. Immunol.* **130,** 663–670.e3 (2012).

156. Tomicić, S. *et al.* High levels of IgG4 antibodies to foods during infancy are associated with tolerance to corresponding foods later in life. *Pediatr. Allergy Immunol.* **20,** 35–41 (2009).

157. Savilahti, E. M. *et al.* Early recovery from cow's milk allergy is associated with decreasing IgE and increasing IgG4 binding to cow's milk epitopes. *J. Allergy Clin. Immunol.* **125,** 1315–1321.e9 (2010).

158. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

159. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32,** 158–68 (2014).

160. Galson, J. D., Pollard, A. J., Trück, J. & Kelly, D. F. Studying the antibody repertoire after vaccination: Practical applications. *Trends Immunol.* **35,** 319–331 (2014).

161. Orton, R. J. *et al.* Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* **16,** 229 (2015).

162. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135,** 183–191 (2012).

163. Prabakaran, P. *et al.* Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. *Front. Microbiol.* **3,** 277 (2012).

164. Bolotin, D. a. *et al.* Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur. J. Immunol.* **42,** 3073–3083 (2012).

165. Baum, P. D., Venturi, V. & Price, D. A. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *Eur. J. Immunol.* **42,** 2834–9 (2012).

166. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30,** 434–9 (2012).

167. Kircher, M., Heyn, P. & Kelso, J. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* **12,** 382 (2011).

168. Michaeli, M., Noga, H., Tabibian-Keissar, H., Barshack, I. & Mehr, R. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front. Immunol.* **3,** 386 (2012).

169. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–85 (1998).

170. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–94 (1998).

171. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–45 (2008).

172. Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* **9,** e1003031 (2013).

173. Larsen, P. A. & Smith, T. P. L. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunol.* **13,** 52 (2012).

174. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–9 (2013).

175. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 9530–9535 (2011).

176. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1347–52 (2012).

177. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14508–13 (2012).

178. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).

179. Ernst, A. *et al.* Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.* **6,** 1782–1790 (2010).

180. Di Niro, R. *et al.* Rapid interactome profiling by massive sequencing. *Nucleic Acids Res.* **38,** e110 (2010).

181. Kretz, C. A. *et al.* Massively parallel enzyme kinetics reveals the substrate recognition landscape of the metalloprotease ADAMTS13. *Proc. Natl. Acad. Sci.* **112,** 201511328 (2015).

182. Ravn, U. *et al.* By-passing in vitro screening - Next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res.* **38,** e193 (2010).

183. Venet, S. *et al.* Transferring the characteristics of naturally occurring and biased antibody repertoires to human antibody libraries by trapping CDRH3 sequences. *PLoS One* **7,** e43471 (2012).

184. Chen, W. *et al.* Characterization of human IgG repertoires in an acute HIV-1 infection. *Exp. Mol. Pathol.* **93,** 399–407 (2012).

185. Ravn, U. *et al.* Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60,** 99–110 (2013).

186. Venet, S., Kosco-Vilbois, M. & Fischer, N. Comparing CDRH3 diversity captured from secondary lymphoid organs for the generation of recombinant human antibodies. *MAbs* **5,** 690–8 (2013).

187. Lomakin, Y. a *et al.* Polyreactive monoclonal autoantibodies in multiple sclerosis: functional selection from phage display library and characterization by deep sequencing analysis. *Acta Naturae* **5,** 94–104 (2013).

188. Lomakin, Y. A. *et al.* Heavy-light chain interrelations of MS-associated immunoglobulins probed by deep sequencing and rational variation. *Mol. Immunol.* **62,** 305–14 (2014).

189. Hu, D. *et al.* Effective Optimization of Antibody Affinity by Phage Display Integrated with High-Throughput DNA Synthesis and Sequencing Technologies. *PLoS One* **10,** e0129125 (2015).

190. Fischer, N. Sequencing antibody repertoires: The next generation. *MAbs* **3,** 17–20 (2011).

191. Chen, H.-S. *et al.* Predominant structural configuration of natural antibody repertoires enables potent antibody responses against protein antigens. *Sci. Rep.* **5,** 12411 (2015).

192. D'Angelo, S. *et al.* Profiling celiac disease antibody repertoire. *Clin. Immunol.* **148,** 99–109 (2013).

193. Ciric, M. *et al.* Metasecretome-selective phage display approach for mining the functional potential of a rumen microbial community. *BMC Genomics* **15,** 356 (2014).

194. Scott, B. M. *et al.* Phage display of the serpin alpha-1 proteinase inhibitor randomized at consecutive residues in the reactive centre loop and biopanned with or without thrombin. *PLoS One* **9,** e84491 (2014).

195. Domina, M. *et al.* Rapid Profiling of the Antigen Regions Recognized by Serum Antibodies Using Massively Parallel Sequencing of Antigen-Specific Libraries. *PLoS One* **9,** e114159 (2014).

196. Dias-Neto, E. *et al.* Next-generation phage display: Integrating and comparing available molecular tools to enable costeffective high-throughput analysis. *PLoS One* **4,** e8338 (2009).

197. Matochko, W. L. *et al.* Uniform amplification of phage display libraries in monodisperse emulsions. *Methods* **58,** 18–27 (2012).

198. 'T Hoen, P. a C. *et al.* Phage display screening without repetitious selection rounds. *Anal. Biochem.* **421,** 622–631 (2012).

199. Ngubane, N. a C. *et al.* High-throughput sequencing enhanced phage display identifies peptides that bind mycobacteria. *PLoS One* **8,** e77844 (2013).

200. Rentero Rebollo, I., Sabisz, M., Baeriswyl, V. & Heinis, C. Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Res.* **42,** e169 (2014).

201. Ryvkin, A. *et al.* Deep panning: Steps towards probing the IgOme. *PLoS One* **7,** e41469 (2012).

202. Liu, X. *et al.* Serum Antibody Repertoire Profiling Using In Silico Antigen Screen. *PLoS One* **8,** e67181 (2013).

203. Heyduk, E. & Heyduk, T. Ribosome display enhanced by next generation sequencing: a tool to identify antibody-specific peptide ligands. *Anal. Biochem.* **464,** 73–82 (2014).

204. Van Blarcom, T. *et al.* Precise and Efficient Antibody Epitope Determination through Library Design, Yeast Display and Next-Generation Sequencing. *J. Mol. Biol.* **427,** 1513–34 (2014).

205. Doolan, K. M. & Colby, D. W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *J. Mol. Biol.* **427,** 328–40 (2015).

206. Olson, C. A. *et al.* Single-round, multiplexed antibody mimetic design through mRNA display. *Angew. Chem. Int. Ed. Engl.* **51,** 12449–53 (2012).

207. Guo, N. *et al.* Reverse Engineering of Vaccine Antigens Using High Throughput Sequencing-enhanced mRNA Display. *EBioMedicine* (2015). doi:10.1016/j.ebiom.2015.06.021

Study overview

Study objectives

In the Introduction, recent developments in high-throughput technologies, specifically HTS coupled with phage display and peptide microarrays, have been outlined. The overall objective of the present studies was to investigate how these technological developments could be applied in epitope mapping. We wanted to determine whether they had advantages over comparable conventional approaches and whether they could provide new biological knowledge. Of note, many of the studies that were mentioned in the introduction have been published concurrent with the present studies. They have provided additional perspective and input as will be debated in the Discussion that follows the four chapters outlined here. It should be emphasized that Chapter 2 and 3 do not represent manuscript drafts and as such are not intended for publication in their current form. They are relatively short research notes that are intended to provide perspective to the main research effort outlined in Chapter 1 and 4.

Chapter 1 - Published manuscript: "High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum"

In the first chapter, it was examined whether combining phage display with HTS could improve the method. There had been very few reports of HTS coupled with phage display and even fewer focused on epitope mapping. The previous reports were troubled by a low number of samples, a lack of control experiments and suggested epitope areas had not been validated. Thus, there was a need to investigate whether the approach could successfully identify epitopes. In order to demonstrate the usefulness of the HTS approach we chose to study sera from peanut allergic patients. These sera were chosen since they were complex samples that would have been difficult to characterize by conventional phage display. Furthermore, the ability to determine epitopes directly in serum could be of clinical interest.

Chapter 2 - "High-throughput sequencing enhanced phage display for discovery of peptidic anti-cobratoxins "

Having observed in Chapter 1 that HTS coupled with phage display provided additional insights into the selected phage pools, we wanted to expand the approach to a different application. Identification of peptide binders is another major application of phage display. In this study, we wanted to examine whether deep sequencing could identify peptide binders that were overlooked in conventional phage display. Through a collaboration with Copenhagen University, we obtained samples that had been analyzed by conventional phage display and sequenced them by HTS. Since the experiments were not performed with HTS analysis in mind, there were some limitations in the experimental design (e.g. samples were not available for all selection rounds and control biopannings applied phages that derived from selections against the target).

Chapter 3 - "Towards epitope mapping of therapeutic candidate antibodies in immunized animal sera"

In Chapter 1, it was demonstrated that phage display coupled with HTS could identify epitopes in sera. In the third study, the goal was to confirm the applicability of the approach in a different context. The experimental setup was similar to Chapter 1; however, the samples were markedly different since they were sera from immunized animals. The hypothesis was that the animals would have relatively high concentrations of relevant antibodies and that the approach could potentially identify several epitopes. In addition, antibodies from the samples were to be thoroughly characterized and the ambition was to correlate knowledge of the epitopes with the additional information.

Chapter 4 – Manuscript in preparation: "High-resolution epitope mapping of peanut allergens demonstrates individualized and highly persistent antibody binding patterns at the residue level"

The epitopes identified in Chapter 1 were validated on a new generation of peptide microarrays. These experiments demonstrated that the microarrays had a great capacity for mapping of linear epitopes. Consequently, we wanted to expand the study to additional sera from peanut allergic patients. The goal

was to use the microarray capacity to characterize peanut epitopes at a resolution that had not previously been achieved. Samples obtained over multiple years were available for some patients and accordingly we wanted to study the developments in epitope patterns over time.

Chapter 1

High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum

Received: 16 March 2015
Accepted: 08 July 2015
Published: 06 August 2015

Anders Christiansen[1], Jens V. Kringelum[2], Christian S. Hansen[2], Katrine L. Bøgh[3], Eric Sullivan[4], Jigar Patel[4], Neil M. Rigby[5], Thomas Eiwegger[6], Zsolt Szépfalusi[6], Federico de Masi[2], Morten Nielsen[2,7], Ole Lund[2] & Martin Dufva[1]

Phage display is a prominent screening technique with a multitude of applications including therapeutic antibody development and mapping of antigen epitopes. In this study, phages were selected based on their interaction with patient serum and exhaustively characterised by high-throughput sequencing. A bioinformatics approach was developed in order to identify peptide motifs of interest based on clustering and contrasting to control samples. Comparison of patient and control samples confirmed a major issue in phage display, namely the selection of unspecific peptides. The potential of the bioinformatic approach was demonstrated by identifying epitopes of a prominent peanut allergen, Ara h 1, in sera from patients with severe peanut allergy. The identified epitopes were confirmed by high-density peptide micro-arrays. The present study demonstrates that high-throughput sequencing can empower phage display by (i) enabling the analysis of complex biological samples, (ii) circumventing the traditional laborious picking and functional testing of individual phage clones and (iii) reducing the number of selection rounds.

The concept of phage display was originally introduced in 1985[1] and is still one of the predominant techniques for the screening of protein-protein interactions. Phage display is based on libraries of phage particles expressing a great variety of exogenous peptides fused to phage surface proteins. The highly diverse library is reduced to a few leads by performing several rounds of selection, also known as bio-panning[2–5]. Phage display has been widely applied in antibody discovery and has given rise to several therapeutic antibodies[6–8]. A range of different types of phage libraries have been constructed, with display of short random peptides or antibody domains as prominent examples[2].

Phage display has been extensively used for the identification of antibody binding sites on antigens, known as epitope mapping[2,9]. Identification of epitopes is an important step in the development of diagnostic tools, in rational vaccine design and in identification of therapeutic targets[10–13]. Epitopes are generally classified as being either linear (continuous) or conformational (discontinuous). Whereas linear epitopes comprise short stretches of consecutive amino acid residues of the primary antigen sequence, conformational epitopes consist of residues that are distant in the primary sequence, but are brought close by the conformational folding of the native protein[9].

[1]Department of Micro- and Nanotechnology, Technical University of Denmark, Kgs. Lyngby, Denmark. [2]Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark. [3]National Food Institute, Technical University of Denmark, Søborg, Denmark. [4]Roche NimbleGen, Madison, Wisconsin, the United States of America. [5]Institute of Food Research, Norwich, United Kingdom. [6]Department of Paediatrics, Medical University of Vienna, Vienna, Austria. [7]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina. Correspondence and requests for materials should be addressed to M.D. (email: martin.dufva@nanotech.dtu.dk)

A prominent issue in phage display is the unintended selection of peptides that do not bind the target of interest, collectively known as "target unrelated peptides" (TUPs)[14,15]. Generally, TUPs either bind a constant part of the screening platform or provide the phage with a proliferation advantage. Constant parts of the screening platform include the solid phase (such as plastic or beads) as well as constant parts of antibodies or other capturing molecules[14,15]. A routine step in phage display is the amplification in bacteria. During this process, a pool of phages propagates in parallel and even minor proliferation advantages can have a profound impact on the output after this step. As a consequence, library diversity is shaped by selection as well as amplification, as thoroughly discussed by Derda *et al.*[16]. One way to deal with the TUP issue is to compare phage clones isolated in an experiment with databases encompassing previously isolated peptides and known TUP motifs[17,18].

Today, the emergence of high-throughput sequencing techniques[19] is transforming multiple areas (e.g. immune repertoire characterization[20]) and is well suited for analysis of the inherent diversity of phage libraries. The data output from deep sequencing should enable the analysis of highly complex samples by phage display and holds the potential to circumvent the traditional laborious picking and functional testing of individual phage clones. Accordingly, in recent years there have been considerable efforts in coupling high-throughput sequencing with phage display. Initially, the potential of this approach was demonstrated by studying protein interactions in a high-throughput manner[21–23] followed by studies of antibody display libraries[24–28]. Recently, a gene-specific library was used to identify immunodominant regions of a vaccine antigen[29]. With regard to random peptide display libraries, an initial study that examined the overlap in peptides obtained by conventional versus high-throughput sequencing coupled phage display[30] has been followed by a few others. One study used deep sequencing to assist the identification of TUP candidates with a propagation advantage[31]. Two other studies employed high-throughput sequencing to identify peptides that bound to cell surfaces[32,33]. However, even though a major application of phage display is in epitope mapping, until now the potential of using high-throughput sequencing enhanced phage display for epitope mapping has only been described to a limited degree. Ryvkin *et al.* demonstrated that a fraction of the peptides obtained after binding to IgG from a pool of HIV positive sera could be aligned to an HIV protein thus indicating some HIV specificity[34]. Liu *et al.* took a broader approach and searched for putative IgG binding targets using BLASTP to determine the similarity between isolated phage peptides and all known proteins[35]. They established the approach on immunised mice as well as a sample from a human melanoma patient. These studies have been very limited with regards to the amount of human patient material and suggested epitopes have not been validated. Nonetheless, these studies encourage further investigations of the potential of high-throughput sequencing-assisted epitope mapping directly on serum.

In this study, we exploit the enormous data output of high-throughput sequencing to investigate samples of great complexity, specifically serum samples from patients with severe peanut allergy. Peanut allergy is regarded as one of the most serious forms of food allergy, in terms of prevalence, persistency and severity[36,37] and Ara h 1, which is the focus of this study, is one of the major peanut allergens[36–38]. We present a general approach to address the issues of TUPs, identify patient-specific epitope motifs directly from serum, and validate identified epitopes using high-density peptide micro-arrays.

Results

Deep sequencing of phage libraries selected against patient serum. The phage selection process (see Fig. 1) was based on the Ph.D.7™ phage library, which displays 7-mer peptides. This library has a reported diversity of 10^9, approaching the theoretical diversity (1.28×10^9), thus covering the majority of possible amino acid combinations. The phage particles were selected over 3 rounds of biopanning against IgE from 12 subject samples, comprising both patients with severe peanut allergy as well as control subjects with no reactivity towards peanut allergens. Specifically, the samples were sera from 4 patients (P1-4) and 4 controls (C1-4) as well as plasma from 2 of these patients (P1 and P4) and 2 of the controls (C2p and C3p). In order to investigate whether there were changes in the epitope patterns over time, the plasma samples from the 2 patients (P1_12 and P4_12) were obtained about 9 years after the serum samples (P1 and P4). Epitope mimicking peptides were selected by competitive addition of Ara h 1. Following PCR amplification the phage DNA was sequenced by high-throughput sequencing. For every one of the 12 samples, DNA was obtained for all 3 selection rounds and subjected to deep sequencing.

For each of the 36 sampled conditions an average of 308,235 (standard deviation (sd) = 102,719) DNA sequences were obtained. Of these, an average of 231,642 (sd = 104,649) could be converted to peptide sequences. Identified peptides were further processed by removing peptides originating from DNA sequences with low quality and not having the NNK codon structure encoded in the phage library. Due to the high prevalence (e.g. > 100,000 copies) of some DNA sequences, derivate sequences (differing in 1 or 2 bases) were frequently observed and removed together with sequences identified as spill-over from one sample to another due to sequencing errors in barcodes (see Materials and Methods as well as Supplementary Fig. S1,S2). Nearly 50% of the initial DNA sequences were discarded in the filtering process resulting in 161,072 (sd = 82,792) peptides per condition on average. To diminish the effect of amplification bias as well as bias towards samples with high number of sequencing reads, the internal rank in each sample, and not the actual sequencing count, for each peptide were used in downstream analyses.

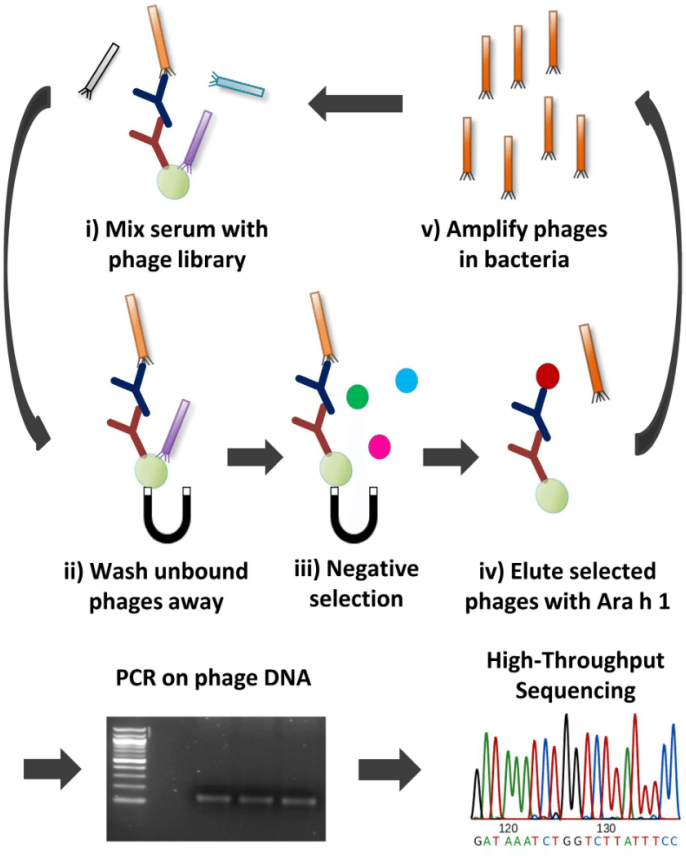


Figure 1. An overview of the phage selection process. (**i**) The IgE from serum is captured by beads coated with anti-IgE. The phage library is added and (**ii**) unbound phages are washed away. (**iii**) Negative selection, specifically incubation with SMP, is carried out to minimise unspecific binding. (**iv**) Phages expressing epitope mimicking peptides are eluted by competitive addition of Ara h 1. (**v**) The eluted phages are amplified in bacteria and either used for another selection round (starting at (**i**) again) or subjected to PCR and high-throughput sequencing.

Examination of the phage selection process. As peptide sequences were obtained for each selection round it was possible to monitor the change in peptide frequencies over the course of successive biopanning rounds. The total number of peptide sequences varied across samples, however, no correlation with selection round was observed (Supplementary Table S2). However, the number of unique peptides in each sample decreased considerably with each subsequent selection round (Fig. 2A and Supplementary Fig. S3). For each sample an average of 89% (sd = 4%) of the peptides identified in round 2 were also present in round 1, excluding peptides with a read count of 1. This was the case for 81% (sd = 9%) of the peptides when comparing round 3 with round 2. These observations indicated that the sequencing depth was adequate to identify the majority of the selected phages.

We also observed that from round 1 to 3 a few peptides had greatly increasing frequencies and ended up dominating each sample (Fig. 2B, Table 1, Supplementary Fig. S3 and Supplementary Table S5). There was a great overlap in the most frequent peptides from round 1 to 3, e.g. for the patient samples an average of 6 (range 4–7) of the ten most frequent peptides from round 1 were also observed after round 3. However, these dominating peptides were often also observed in control samples indicating that they were TUPs with a non-specific selection advantage. In fact, on average 6 out of the 10 most frequent peptides in each selection round were also observed in the control samples (Supplementary Table S5). However, the general overlap in peptides between patient and control samples was much smaller (Fig. 3A). This indicates that the highly prevalent peptides are more likely to emerge in multiple samples,

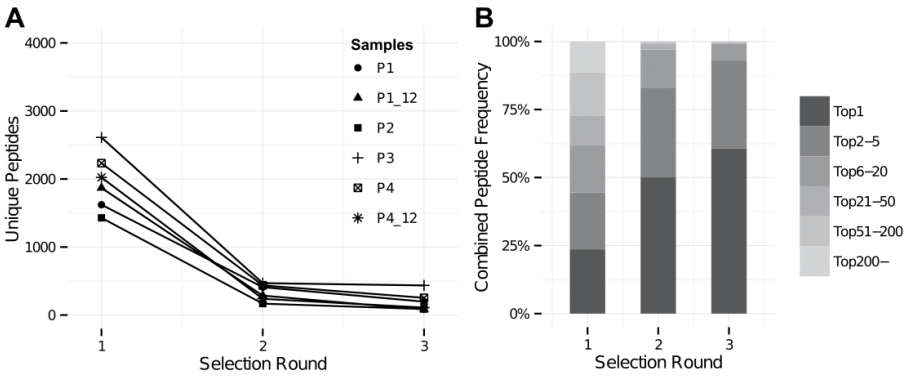


Figure 2. **(A)** A plot of the number of unique peptides identified in each patient sample for selection round 1–3. In the first selection round an average of 1827 (sd = 642) unique peptides were obtained, dropping to 313 (sd = 150) and finally 170 (sd = 111) unique peptides after the second and third selection round, respectively. Results were similar for control samples as shown in Supplementary Fig. S3. **(B)** A stacked bar chart showing the combined frequency of certain rank intervals. Specifically, the frequency of the most frequent peptide (Top1) along with the combined frequencies of the peptides ranked 2–5, 6–20, 21–50, 51–200 and below 200. The average of the combined frequencies for all patients is shown. Results were similar for control samples as shown in Supplementary Fig. S3.

Round 1		Round 2		Round 3	
WSLSELH	14.3%	VSRDTPQ	33.9%	VSRDTPQ	73.1%
DGRYYIN	9.1%	DGRYYIN	20.2%	DGRYYIN	9.5%
VSRDTPQ	4.5%	WSLSELH	15.8%	WSLSELH	8.7%
MHTVAVQ	4.2%	HWRLPLH	7.4%	HLNQQNH	3.4%
HLNQQNH	3.5%	HLNQQNH	4.6%	HWRLPLH	0.9%
APRNVPP	2.4%	IWRLPTH	4.2%	IWRLPTH	0.7%
WPFHGDN	1.5%	MHTVAVQ	2.4%	SYTDLLR	0.7%
QISAASQ	1.5%	VMPLDDV	1.4%	VMPLDDV	0.5%
LPYDHLP	1.5%	SYTDLLR	1.2%	YPWFIRA	0.5%
YPWFIRA	1.4%	YPWFIRA	0.9%	MHTVAVQ	0.3%

Table 1. The ten most frequent peptides and their frequencies for a sample patient, P1. The peptides are separated according to selection round (Round 1–3). The peptides with a grey overlay were also observed in a control sample with a prevalence >1. Similar top 10 tables for the other samples have been included in Supplementary Table S5.

and hence be classified as TUPs. The most widespread peptide was the QLYREFN peptide, which was identified in 11 out of 12 samples. It is presented in Supplementary Table S6 alongside other TUP candidates to guide future phage display investigations.

Next, to extend the comparison of patient and control peptides beyond exact matches peptides were classified based on their similarity to other peptides. This was accomplished by pairwise alignment. The alignment threshold was established by doing 10 million random peptide comparisons, and using the 99.999 percentile alignment score. Based on this threshold, a comparison of peptide similarity between patients and controls was made (Fig. 3B). A much higher degree of overlap was observed compared to the observations based on exact matches. This indicates that peptides with sequence similarity were frequent, and that similarity had to be taken into account in further analyses.

Identifying patient-specific peptide clusters. The next step was based on the hypothesis that selection of related peptides in multiple patients would represent key selection motifs that would be the main Ara h 1 epitope candidates. Peptides were clustered by pairwise alignment (Fig. 4). A p-value was derived by permutation of patient and control labels using a test statistic based on whether the peptides populating the cluster were derived from a patient sample or a control sample as well as their internal

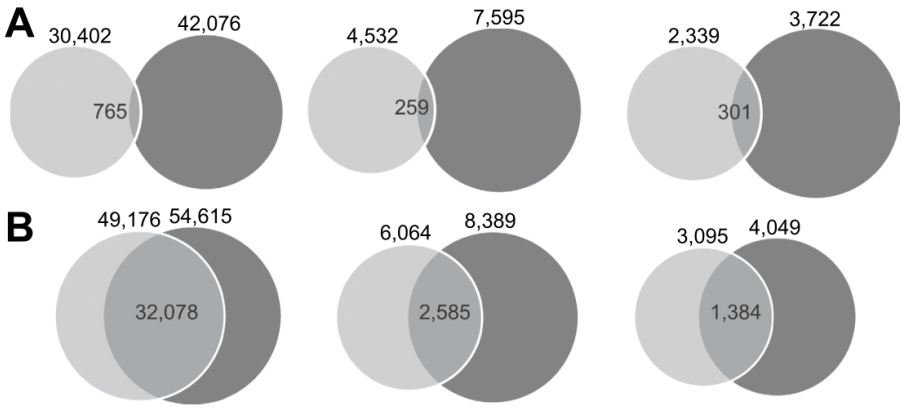


Figure 3. **(A)** Venn diagrams showing the overlap in identical, unique peptides between patient samples (light circle) and controls (dark circle) for each selection round. The exact numbers have been included in the figure. **(B)** Venn diagrams showing the overlap in similar peptides between patient samples (light circle) and controls (dark circle) for each selection round. Similarity was determined by pairwise alignment using an alignment threshold established as the 99.999 percentile alignment score for 10 million random peptide comparisons. The specific numbers of peptides in each group have been included in the figure.

rank score. Accordingly, clusters that have multiple patient-specific peptides with a high rank, and where similar peptides could not be obtained in the control samples, would be provided with the lowest p-value. As such, the clustering approach takes the observed similarity between peptides from patient and control samples (Fig. 3B) into account. Interestingly, after statistical correction for the multiple testing problem (by Dunn-Bonferroni correction), a single statistically significant cluster stood out (p-value < 0.01) in selection round 2 and 3 (highlighted in Fig. 4A). In round 1 a patient-enriched cluster also had a low p-value, however, there were a large number of clusters in this round and the p-value was no longer significant after correction for multiple testing. The number of peptides in the significant cluster were 28 in selection round 2 and 17 in round 3 (Supplementary Table S7). The patient samples contributed a different number of peptides to the significant cluster (Supplementary Table S8). However, the patient sample P1 had most peptides in each of the clusters, specifically 46% and 47% of the peptides in the second and third selection round, respectively. P1, P3 and P4 contributed peptides to the significant clusters in both selection round 2 and 3. P1 and P4, which were investigated at two time points, contributed peptides at both time points. On the contrary, no peptides from P2 populated the significant clusters.

To examine which amino acid residues were the most conserved in the significant clusters, a sequence logo plot was made (Fig. 4B,C). Overall, the peptide motifs that appeared were highly similar. Closer inspection revealed that the amino acids W, P and H appeared particularly important.

Identifying epitope candidates. In order to investigate whether the selected peptide cluster matched the primary Ara h 1 protein sequence, thus suggesting a linear epitope, all peptides were aligned to the primary sequence of Ara h 1 (Fig. 5). Multiple patient-derived peptides aligned well around position 136. This was not the case for peptides derived from control samples as well as randomly generated peptides (Fig. 5). The amino acid sequence around this position was WRRPSH. The majority of the peptides with prominent alignment scores were part of the significant cluster identified in selection round 2 and 3 in Fig. 4.

Epitope validation on peptide arrays. To investigate the authenticity of the epitope suggested by alignment (Fig. 5), we investigated serum reactivity towards Ara h 1 on peptide micro-arrays. Overlapping 12-mer peptides with a single offset covering the Ara h 1 sequence were synthesised in triplicates on the chip (a total of 1845 peptides). Binding of serum IgE to the peptides was then investigated (Fig. 6). Three out of four patients (P1, P3 and P4) showed distinct peaks of IgE reactivity at position 136 in Ara h 1, which is also where the peptides from the significant cluster aligned. These three patients were the same three patients that had peptides that aligned well to Ara h 1 at position 136 (Fig. 5) and populated the significant cluster (Fig. 4A). As such, the epitope suggested by clustering and alignment was confirmed by the peptide array observations. Both time points investigated for P1 and P4 demonstrated reactivity to the same epitope thus confirming the persistence suggested by the phage peptides. For the P2 patient no IgE reactivity was observed on the peptide array, and this patient did not contribute any peptides to the significant clusters. Similarly, no IgE reactivity was observed for control subjects (Supplementary Fig. S4).

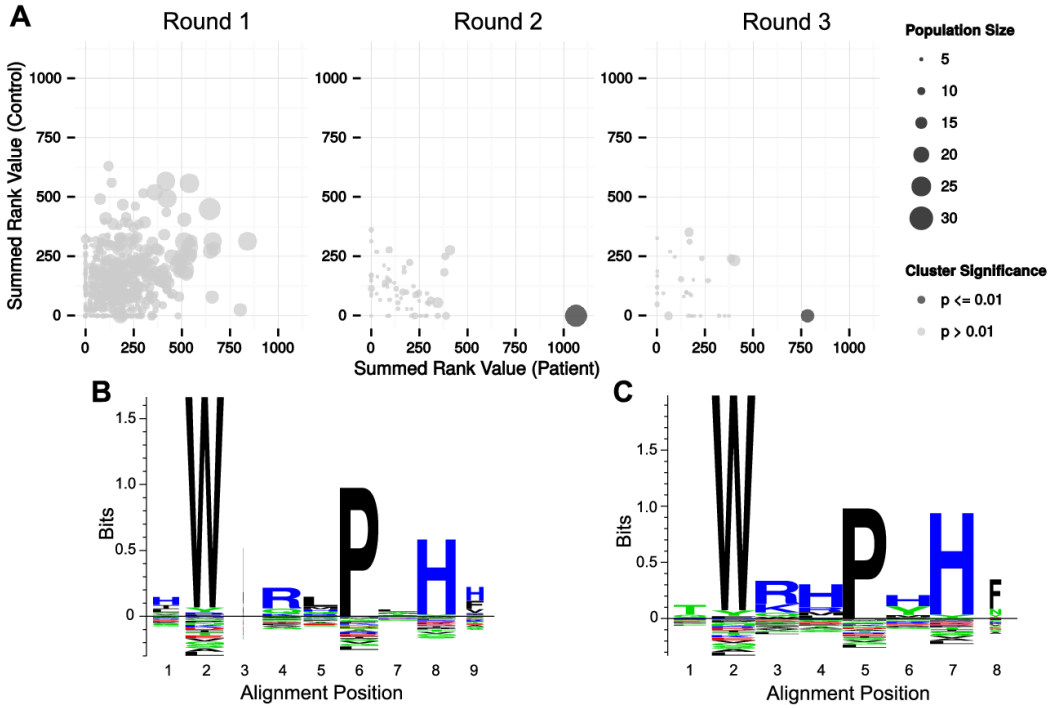


Figure 4. **(A)** Plots of the summed rank value for clusters obtained by pairwise-alignment and separated according to selection round (1–3). The clusters are further separated based on whether the summed rank value derives from peptides identified in patients (x-axis) or controls (y-axis). 4624, 617 and 333 clusters were identified in selection round 1, 2 and 3, respectively. One cluster in selection round 2 and 3 was significantly enriched (p-value < 0.01) for patient-specific peptides and is coloured dark. **(B)** Peptide sequence logo plot of the significant cluster obtained in round 2 and highlighted in (**A**). The logo was made by constructing position-weighted Kullback Leibler logos of the multiple alignment of the cluster peptides, using the Seq2Logo webserver[51]. Height corresponds to the amount of information contained in a specific position. Large symbols represent frequently observed amino acids. Symbols are narrow if there were many gaps in the alignment at the position. The Seq2Logo default amino acid colour coding is used (DE residues are red, NQSGTY are green, RKH are blue and the remaining are black). **(C)** Similar to (**B**) but for the significant cluster in round 3.

Following the confirmation of the epitope authenticity the frequency of the peptides that enabled the epitope identification were re-examined. Overall, the peptides had low frequencies, which would have made them very difficult to detect by traditional Sanger sequencing of phage clones (Supplementary Fig. S5). Assuming that 1000 phage clones were screened, e.g. by ELISA, thereby detecting peptides with a frequency above 0.1%, only 6 and 7 out of 28 and 17 peptides in the significant clusters would have been detected in selection round 2 and 3, respectively. As a consequence, the identified epitope would have been missed for P4 in all selection rounds and also for P3 in the second selection round.

Discussion

Harnessing the power of high-throughput sequencing can improve phage display in a number of ways. First of all, the extensive data output enables the detection of rare phages in a larger population. This feature was used to detect antibody epitopes in polyclonal serum samples. It is worth noting that the IgE antibodies that target Ara h 1 only comprise a tiny fraction of the serum antibodies, given that IgE is a rare isotype compared to e.g. IgG[39] and, further, only a fraction of the IgE consists of peanut-specific antibodies[40]. Nonetheless, Ara h 1 epitopes could be identified in the patients.

The bioinformatic pipeline presented here, which is based on clustering and contrasting with control samples, assisted the identification of a relevant peptide motif. It is worth emphasising that the clustering approach is very general, and multiple selection motifs could, in principle, be identified simultaneously. As such, the pipeline should be widely applicable in the analysis of multi-target mixtures in a variety of settings extending beyond epitope mapping to e.g. binding assays and cell surface analysis. Furthermore,

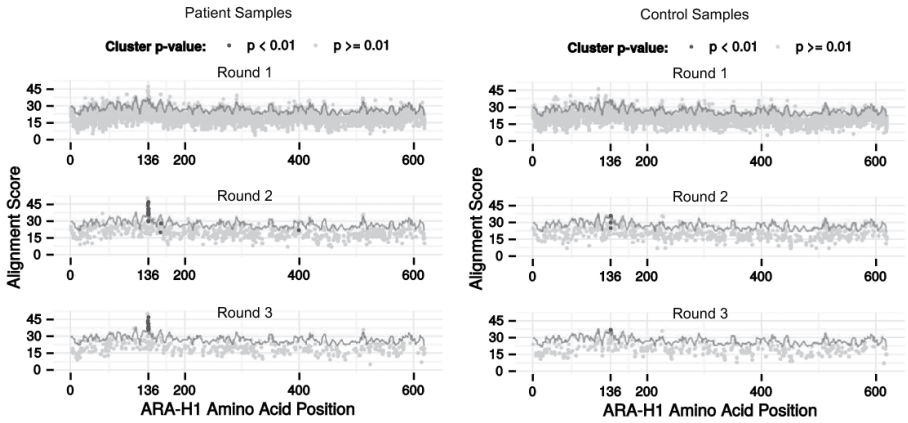


Figure 5. Alignment of patient-derived peptides (left) or control-derived peptides (right) to the primary sequence of Ara h 1. Peptides were mapped to Ara h 1 by aligning each peptide to overlapping 7-mers from the Ara h 1 primary sequence using pairwise alignment. Each phage peptide was plotted at the starting position of the overlapping 7-mer that provided the best alignment score. The top row is a plot for the peptides from selection round 1, while the middle and bottom row are for selection round 2 and 3, respectively. Peptides that were part of a statistically significant cluster are emphasised as dark spots. The grey line represents a significance assessment which is the 99.99 percentile alignment scores obtained after aligning 1 million random peptides to each overlapping 7-mer. Alignments with a score below zero were discarded.

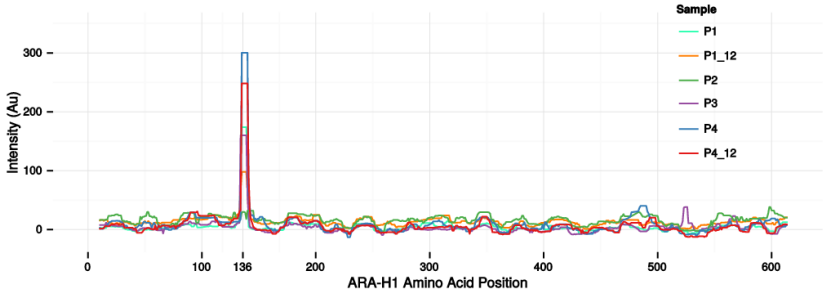


Figure 6. IgE reactivity of patients to Ara h 1as measured by peptide micro-arrays. The right-bound rolling median (window size 12) of the mean intensity of the triplicate 12-mer peptides overlapping each residue is shown for every patient sample. The start position of the identified epitope, position 136, has been specifically marked. For P2 no peaks were observed, whereas all the other samples showed a distinct peak at position 136.

with regards to epitope mapping, partly due to the high diversity of the phage library (approximately 10^9), the approach is not limited to mapping of linear epitopes. If a highly patient-specific motif emerged, that did not match the primary sequence of the target of interest, it could be a selection motif that mimics a conformational epitope.

Another promising observation in this study, is that a patient-specific motif could already be observed in early selection rounds, supporting the notion that deep sequencing will enable phage display experiments with fewer, and ultimately, a single selection round. By performing fewer selection rounds the selection bias can be reduced which should diminish the loss of relevant sequences[16,31,32]. Hoen *et al.* have previously suggested the idea, however, they identified phages of interest based on persistence across selection rounds[32]. Similarly, other studies have identified peptides of interest based on enrichment through consequtive biopanning rounds[33,41]. These approaches are obviously only possible if multiple selection rounds are sequenced or, alternatively, a thorough (and expensive) sequencing of the naïve library could be sufficient. Motif identification represents an alternative approach, that could enable the identification of peptides of interest based on sequencing a single round.

Interestingly, in this study the epitope was identified without the functional testing which is commonly part of traditional phage display. The colony picking, functional testing (e.g. by ELISA) and subsequent Sanger sequencing of positive clones is a very laborius element in phage display[9,42]. By omitting the functional step and performing fewer selection rounds, we suggest that the throughput of phage display can be increased, as multiple complex samples can be tested in parallel.

Deep sequencing of the phage output after each selection round enabled a thorough examination of the selection process. The general trends were highly consistent across samples. With each round fewer unique peptides were obtained; decreasing from thousands after the first round to a few hundred after three rounds. This observation is in line with the expected process of affinity selection, wherein the specificity of the phage-libraries increases with additional biopanning rounds while the diversity drops[16,43].

There was a great overlap among the most frequent peptides across selection rounds. However, recurring peptides with high frequencies could also be observed in control samples indicating that they were TUPs. This finding suggests that consistently prevalent peptides identified in other studies[32,34] might be TUPs, which would explain why they did not match the target. Along these lines, many phage experiments have likely been discarded, or demanded laborious optimisation, because the only hits identified were irrelevant. The prevalence of TUPs was further suggested by the observation that the peptide overlap between paired samples obtained at two time points were similar to the overlap to other patient samples (for the paired samples the median Tanimoto coefficient for unique peptides was 0.6% and 0.5% for the matching pair and 0.8% and 0.4% for the overlap to the remaining patient samples for P1 and P4, respectively).

Derda *et al.* recently identified hundreds of peptides with proliferation advantages, so-called 'parasitic' TUP peptides, in the Ph.D.-7™ library[31]. We did not observe any overlap between the peptides we obtained, and the reported parasitic TUPs. We speculate that selection criteria and experimental context (such as library type and batch) vary greatly between phage experiments and have an extensive impact on which TUPs will be selected. The high number of TUPs we observed, most of which were not previously reported, demonstrates how critical it is to filter out these sequences. As a result, we suggest including parallel control experiments to identify putative TUPs in future experimental setups. Until TUP databases[17,18] have become more comprehensive, it seems to be the optimal way to take into account the distinct differences between selection schemes. It is worth noting that careful experimental design could allow parallel analysis of two or more conditions of interest, with one acting as control for the others and vice versa. This may not be feasible for all setups, but would omit the need to include a condition, that only provides information on TUPs.

Possibly due to the sample complexity a great variety of peptides were observed in the output. Selection of related peptides across patients represented key selection motifs that assisted the epitope identification. The alignment threshold used in the clustering process was determined based on simulations with randomly generated peptides. The threshold was relatively conservative and could have been varied to increase or decrease alignment stringency. A consequence of the conservative threshold is that relevant sequences may be missed. The importance of the threshold level is underlined by the observation that some patient samples had very few peptides in the significant cluster and if the alignment threshold had been slightly more restrictive, the epitope might have been overlooked. The fact that we observe peptides of high similarity that can be clustered and form motifs underlines the importance of taking similar peptides into account instead of focusing only on identical matches. Importantly, this also applies for TUPs, where excluding peptides with exact matches in control samples, such as it has previously been done[35], may not be sufficient. The bioinformatic approach we have presented solves this issue, since TUP sequences from patients would cluster with similar peptides obtained in control samples. Consequently, even when there are no exact matches between a patient peptide and a control peptide, clustering of patient peptides with similar control peptides would result in a mixed cluster that would be disregarded due to its p-value. Alternative clustering approaches are available for the identification of peptide motifs in large-scale data, such as MUSI[44]. However, MUSI does not contrast the results with control samples and, furthermore, duplicate copies of the same sequence are removed. We suggest that the prevalence of the peptides has some correlation with the importance of the peptides and, accordingly, we employ a rank score in our analyses.

Deep sequencing of phage libraries introduced some inherent issues that had to be addressed. Highly prevalent DNA sequences gave rise to derivative sequences due to sequencing errors (Supplementary Fig. S1) as previously reported by Matochko *et al.*[31]. We further observed that, as samples were multiplexed for deep sequencing, highly prevalent DNA sequences consistently turned up with low counts in other samples on the same sequencing chip (Supplementary Fig. S2). This is presumably due to errors in sequencing of the sample barcode. It is of vital importance for later alignment and clustering that derivative sequences (both internal 'mutations' and cross-sample 'spill-over') are eliminated. Otherwise, peptides could be assigned to the wrong sample, which could hinder TUP identification, and highly prevalent peptides would give rise to clusters, that were derived from artificial sequences. The issue with derivative sequences, which is inherent to any deep sequencing study where some DNA sequences have a very high prevalence, would clearly benefit from studies dedicated to describing and solving these issues.

There was an extensive overlap in the epitope mapping results for individual samples from phage display compared to peptide micro-arrays. Specifically, the samples that had peptides mimicking the Ara h 1 epitope were also the samples where this epitope could be identified on the peptide micro-array. This

suggests that both techniques generate reproducible results of considerable accuracy. Interestingly, the identified epitope appeared to be persistent, since it could be identified by phage display as well as peptide micro-arrays in samples obtained with a 9 year interval. Of note, the identified epitope is included in a previously described epitope hotspot of Ara h 1[36,45]. A previous study, based on conventional phage display, further suggested conformational Ara h 1 epitopes[46], however, no such motifs were identified in the present study. For patient P2 no epitopes were identified by phage display or peptide micro-arrays. This patient had a low level of specific IgE and we speculate that the methods were not sensitive enough to identify any epitopes or that the sample quality was low.

The coupling of phage display with high-throughput sequencing offers several advances to conventional phage display. First of all, the large data output provides an opportunity to perform in-depth studies of the phage selection process and also permits the identification of rare phages in a large population. This should improve knowledge of the parameters that affect phage selection and enable the study of multi-target samples. Here, a general bioinformatic approach has been presented, which identifies selection motifs of interest. The observed prevalence of TUPs demonstrated the necessity of examining parallel control samples, which were used as a contrast in the clustering approach. Relevant selection motifs could be identified in early selection rounds and without laborious functional screening. The increased throughput combined with the in-depth data output holds the promise to expand the use of phage display to new applications.

Methods

Peanut allergic patient samples. Sera from 4 peanut allergic adults (P1-4) and 4 peanut-tolerant control subjects (C1-4) were analysed alongside plasma from 4 of these individuals (P1_12, P4_12, C2p and C3p). The 2 patient plasma samples (P1_12 and P4_12) were obtained about 9 years after the patient sera samples. The 4 patients had a convincing history of peanut-related anaphylaxis according to WAO guidelines[47]. All samples were tested for peanut extract-specific IgE by ImmunoCAP and Ara h 1-specific IgE by ELISA. No Ara h 1-IgE reactivity was observed for the control samples. The protocol was approved by the local Ethical Committee of the Medical University of Vienna and written informed consent was obtained (protocol no. EK 428/2008). The samples were treated in accordance with the approved guidelines.

Selection of Ara h 1 mimicking peptides. Figure 1 provides an overview of the selection process. In brief, $12 \times 200\,\mu L$ (6 mg) of M-280 tosylactivated Dynabeads (Invitrogen) were coated with $10\,\mu g$ polyclonal rabbit anti-human IgE (DAKO, A0094) for 48 h at 37 °C, and blocked with 0.5% skimmed milk powder (SMP) in Phosphate Buffer Saline (PBS). Patient serum ($500\,\mu L$) or plasma ($1{,}300\,\mu L$) was diluted in PBS with 0.5% SMP and 0.05% Tween 20, and incubated overnight at 4 °C with $200\,\mu L$ of the coated bead suspension. After extensive washing, $50\,\mu L$ of the coated beads were used immediately for the first round of selection, while the remaining $150\,\mu L$ were stored for later selection rounds.

In the first selection round $50\,\mu L$ beads with immobilized IgE were blocked in 1 mL of 2% SMP in PBS for 1 h at room temperature and washed extensively. The beads were then incubated overnight at 4 °C with $10\,\mu L$ ($\sim 2 \times 10^{11}$) phages from a library of phages displaying random heptamer peptides (Ph.D.-7™, New England BioLabs). This was followed by extensive washing and negative selection; specifically 1 h incubation in 10 mL PBS with 0.5% SMP and 0.05% Tween 20 and 8 washing steps each consisting of 10 min incubation in 10 mL PBS with 0.05% Tween 20. Next, phages of interest were eluted by adding $25\,\mu g$ Ara h 1 protein. The Ara h 1 was purified as previously described by Bøgh *et al.*[48]. The purity of Ara h 1 was confirmed by reverse phase high-performance liquid chromatography to be >99%. Size exclusion chromatography demonstrated that the Ara h 1 was present as a trimer. Eluted phages were amplified by direct infection of ER2738 Escherichia coli cells (New England Biolabs). Phage amplification was done at 37 °C for 4.5 hours. Amplified phages were precipitated using PEG/NaCl and titered. The second and third selection rounds were carried out in an identical way, using $\sim 2 \times 10^{11}$ of the phages selected in the previous round.

High-throughput Sequencing. Amplified phage eluates from each of the 3 selection rounds were subjected to PCR using the Phusion High-Fidelity DNA polymerase (New England Biolabs). To mitigate PCR bias, the PCR reactions were carried out in triplicates and the number of cycles was limited to 25. The PCR was carried out as recommended by the manufacturer, using $25\,\mu L$ reactions with $1\,\mu L$ 1:10 diluted phage eluates as template and an annealing temperature of 55 °C. The variable part of the phage library was specifically amplified using barcoded primers (Supplementary Table S1). The corresponding PCR fragment was extracted from a 2% agarose gel using the Qiaquick Gel Extraction Kit (Qiagen). The triplicate PCR reactions were pooled prior to the gel extractions. The DNA concentration was determined using a Qubit fluorometer (Life) and the fragment size was confirmed using a 2100 Bioanalyzer (Agilent) following manufacturer's protocol.

Next, sequencing adaptors were ligated onto the DNA fragments using the Ion Plus fragment library kit (Life), and the fragment size and concentration was confirmed using a Bioanalyzer (Agilent) and Qubit fluorometer (Life), respectively. Subsequently, the DNA fragments were attached to Ion Sphere Particles (ISPs) and amplified by emulsion PCR according to the Ion OneTouch 200 protocol. The ISPs

were loaded onto an Ion 316 chip and sequenced on an Ion Torrent PGM using the Ion Torrent sequencing kit 200 v2, as described by the manufacturer.

Converting sequencing reads to peptides and quality filtering. DNA reads obtained from the sequencing was first filtered by sample barcodes. Reads that did not match any of the barcodes, both forward and reverse complement, were removed. The DNA reads were then translated in all six reading frames and aligned to the protein regions of the phage protein that flank the displayed 7-mer peptide using the BLAST algorithm[49]. Then 7-mer peptides and their originating DNA sequences were extracted from reads that produced meaningful alignments; i.e. the protein sequence aligned to both flanking regions with a 7-mer peptide gap.

The peptides were further processed by i) removing peptides where at least 1 DNA base pair (bp) coding for the peptide were assigned a Phred quality score below 20 and ii) removing reads where the DNA sequence did not match the NNK-codon pattern intrinsic to the Ph.D.-7™ phage library (N = any nucleotide, K = G or T).

Some sequences had a very high prevalence within a sample. A consequence of high-throughput sequencing read errors[19,31] could be that such abundant sequences might artificially give rise to similar sequences with a low abundance, differing from the abundant sequence at 1 or 2 positions. In fact, a linear relationship was observed between the prevalence of abundant sequences and the prevalence of similar "derivative" sequences (Supplementary Fig. S1). This suggests that the derivative peptides are indeed read error variations of the abundant peptide. As a result, sequences that were at least 500 or 10,000 times less prevalent than a similar (high abundant) sequence in the same sample, differing by just one or two base pairs, respectively, were removed. These conservative thresholds were established to ensure the relevance of the detected peptides. Furthermore, when two subjects were analysed on the same Ion Torrent chip, highly abundant sequences in one patient were sometimes observed with a low prevalence in the other subject. The relationship in prevalence was, once again, linear and such relationship could not be observed across different sequencing chips (Supplementary Fig. S2). The derivative sequences likely originate from sequencing errors in the barcode sequence leading to a mislabelling of the sample. To avoid this issue, peptides of low abundance were removed if they were also observed with at least 100 times higher prevalence in the other subject tested on the same Ion Torrent chip. Finally, the obtained peptides were compared to a list of "parasitic" phages with a suspected proliferation advantage[31], however, no overlap was observed and no peptides were removed. An overview of the number of reads that passed each step of quality filtering can be found in Supplementary Table S2-4.

Peptide clustering. To remove bias towards samples with a higher number of sequencing reads, each unique peptide in a sample was assigned a normalised rank score based on the following formula:

$$r_i = 100 - \frac{R_i}{N} \times 100 \tag{1}$$

Here R_i is the rank of peptide i within a specific sample. Peptides with the same prevalence were given an average rank score. As a result of the above conversion, the most prevalent peptide in each sample will be given a rank score of 100, and the least prevalent peptide will have a rank score of 0. The rank values were summed across samples and used for sorting peptides prior to clustering.

Peptides of each sample were clustered according to the Hobohm I algorithm using pairwise alignment as distance function[50] and an alignment score threshold of 31. The alignment threshold of 31 was established as the 99.999 percentile of 10 million pair-wise alignments of random peptides, using a standard Smith-Waterman alignment algorithm (using BLOSUM50, gap-opening and extension penalties of 5 and 1, respectively, and a minimum alignment length of 6). This alignment threshold was also used for an overall comparison of the peptides in patient and control samples.

To establish patient-related significance each cluster was analysed by patient-control contrasting. The difference, t, between summed rank-scores, r_i, assigned to patients and controls within a single cluster were calculated and used as test-statistic. The null-hypothesis assumes that the test statistic would be equal or higher to that of random clusters established by permutation, i.e. shuffling the patient-control labels 10,000 times. A p-value of the observed test-statistic was calculated as the number of random clusters having a lower test-statistic divided by the number of permutations (10,000), as shown in Equation 2:

$$p = \frac{N\left(t_{true} < t_{permutated}\right)}{N} \tag{2}$$

The null-hypothesis was rejected for clusters with a p-value lower than $0.01/n$ where n is the number of clusters formed. Hence, such clusters had a significant patient-control contrast and were assumed to be obtained from allergen-specific selection.

Representing significant clusters by sequence motifs. Cluster motif visualisation was done by constructing position-weighted Kullback Leibler logos of the multiple alignment of the cluster peptides,

using the Seq2Logo webserver[51]. The multiple alignment was performed using a MAFFT-incorporating Needleman-Wunsch global pairwise alignment algorithm using 1,000 cycles of iterative refinement for increased accuracy.

Aligning peptides to the Ara h 1 sequence. Peptides that were selected from the phage libraries were mapped to Ara h 1 by aligning each peptide to overlapping 7-mers from the Ara h 1 primary sequence (Uni-prot ID: P43238) using the same alignment parameters as for the peptide clustering. Each phage peptide was assigned the starting position of the overlapping 7-mer that provided the best alignment score. Alignments with a score below zero were discarded. The significance of alignment scores was assessed by aligning 1 million random peptides to each overlapping 7-mer and reporting the 99.99 percentile of the alignment scores.

High-density peptide micro-arrays. A peptide library was generated *in silico* for synthesis on high-density peptide micro-arrays. The library consisted of single-offset overlapping 12-mer peptides covering the primary sequence of Ara h 1 (Uni-prot ID: P43238) in triplicates (1845 peptides). Peptide synthesis was accomplished by light-directed array synthesis in a Roche-NimbleGen Maskless Array Synthesizer (MAS) using an amino functionalised substrate coupled with 6-amino hexanoic acid as a spacer and amino acid derivatives carrying a photosensitive 2-(2-Nitrophenyl) propyl-oxy-carbonyl-group (NPPOC). Coupling of amino acids was done using pre-activated amino acid with activator (HOBT/HBTU) and Ethyl-di-iso-propylamine in DMF for 5–7 minutes before flushing the substrate. Cycles of coupling were repeated until 12-mer peptides were synthesized. Intermediate washes on the arrays were done with N-Methyl-2-pyrrolidone (NMP) and site-specific cleavage of the NPPOC group was accomplished by irradiation of an image created by a Digital Micro-Mirror Device (Texas Instruments, SXGA + graphics format), projecting light with a 365 nm wavelength. Final de-protection to cleave off the side-chain protecting groups of the amino acids was done with Trifluoroacetic acid(TFA)/Water/Triisopropylsilane for 30 minutes. The final micro-arrays were incubated overnight at 4 °C with individual sera mixed with binding buffer in a 1:5 dilution in a final volume of 20 μL. This was followed by incubation with goat anti-Human IgE conjugated with tetramethylrhodamine (TRITC) (Life, A18798) at room temperature for 3 hours. Serum from P1 and P3 were also evaluated in a 1:2 dilutions with binding buffer. Finally, the arrays were scanned using a MS200 microarray scanner, and signals were extracted using NimbleGen DEVA signal extraction software.

References

1. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228,** 1315–7 (1985).
2. Pande, J., Szewczyk, M. & Grover, A. Phage display: concept, innovations, applications and future. *Biotechnol. Adv.* **28,** 849–58 (2010).
3. Feldhaus, M. J. & Siegel, R. W. Yeast display of antibody fragments: A discovery and characterization platform. *J. Immunol. Methods* **290,** 69–80 (2004).
4. Bessette, P. H., Rice, J. J. & Daugherty, P. S. Rapid isolation of high-affinity protein binding peptides using bacterial display. *Protein Eng. Des. Sel.* **17,** 731–739 (2004).
5. Zahnd, C., Amstutz, P. & Plückthun, A. Ribosome display: selecting and evolving proteins *in vitro* that specifically bind to a target. *Nat. Methods* **4,** 269–279 (2007).
6. Hoogenboom, H. R. Overview of antibody phage-display technology and its applications. *Methods Mol. Biol.* **178,** 1–37 (2002).
7. Nelson, A. L., Dhimolea, E. & Reichert, J. M. Development trends for human monoclonal antibody therapeutics. *Nat. Rev. Drug Discov.* **9,** 767–74 (2010).
8. Bradbury, A. R. M., Sidhu, S., Dübel, S. & McCafferty, J. Beyond natural antibodies: the power of *in vitro* display technologies. *Nat. Biotechnol.* **29,** 245–54 (2011).
9. Rowley, M. J., O'Connor, K. & Wijeyewickrema, L. Phage display for epitope determination: A paradigm for identifying receptor-ligand interactions. *Biotechnol. Annu. Rev.* **10,** 151–188 (2004).
10. Irving, M. B., Pan, O. & Scott, J. K. Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics. *Curr. Opin. Chem. Biol.* **5,** 314–324 (2001).
11. Wang, L.-F. & Yu, M. Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Curr. Drug Targets* **5,** 1–15 (2004).
12. Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Freund, N. T. & Weiss, Y. Epitope mapping: The first step in developing epitope-based vaccines. *BioDrugs* **21,** 145–156 (2007).
13. Nagata, S. & Pastan, I. Removal of B cell epitopes as a practical approach for reducing the immunogenicity of foreign protein-based therapeutics. *Adv. Drug Deliv. Rev.* **61,** 977–985 (2009).
14. Menendez, A. & Scott, J. K. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. *Anal. Biochem.* **336,** 145–57 (2005).
15. Vodnik, M., Zager, U., Strukelj, B. & Lunder, M. Phage display: Selecting straws instead of a needle from a haystack. *Molecules* **16,** 790–817 (2011).
16. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16,** 1776–1803 (2011).
17. Huang, J., Ru, B., Li, S., Lin, H. & Guo, F. B. SAROTUP: Scanner and reporter of target-unrelated peptides. *J. Biomed. Biotechnol.* **2010,** 101932 (2010).
18. Huang, J. *et al.* MimoDB 2.0: A mimotope database and beyond. *Nucleic Acids Res.* **40,** D271–7 (2012).
19. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
20. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32,** 158–68 (2014).
21. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).

22. Ernst, A. *et al.* Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.* **6,** 1782–1790 (2010).
23. Di Niro, R. *et al.* Rapid interactome profiling by massive sequencing. *Nucleic Acids Res.* **38,** e110 (2010).
24. Ravn, U. *et al.* By-passing *in vitro* screening—Next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res.* **38,** e193 (2010).
25. Venet, S. *et al.* Transferring the characteristics of naturally occurring and biased antibody repertoires to human antibody libraries by trapping CDRH3 sequences. *PLoS One* **7,** e43471 (2012).
26. Chen, W. *et al.* Characterization of human IgG repertoires in an acute HIV-1 infection. *Exp. Mol. Pathol.* **93,** 399–407 (2012).
27. Ravn, U. *et al.* Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60,** 99–110 (2013).
28. Venet, S., Kosco-Vilbois, M. & Fischer, N. Comparing CDRH3 diversity captured from secondary lymphoid organs for the generation of recombinant human antibodies. *MAbs* **5,** 690–8 (2013).
29. Domina, M. *et al.* Rapid Profiling of the Antigen Regions Recognized by Serum Antibodies Using Massively Parallel Sequencing of Antigen-Specific Libraries. *PLoS One* **9,** e114159 (2014).
30. Dias-Neto, E. *et al.* Next-generation phage display: Integrating and comparing available molecular tools to enable costeffective high-throughput analysis. *PLoS One* **4,** e8338 (2009).
31. Matochko, W. L., Cory Li, S., Tang, S. K. Y. & Derda, R. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res.* **42,** 1784–98 (2014).
32. 'T Hoen, P. A. C. *et al.* Phage display screening without repetitious selection rounds. *Anal. Biochem.* **421,** 622–631 (2012).
33. Ngubane, N. A. C. *et al.* High-throughput sequencing enhanced phage display identifies peptides that bind mycobacteria. *PLoS One* **8,** e77844 (2013).
34. Ryvkin, A. *et al.* Deep panning: Steps towards probing the IgOme. *PLoS One* **7,** e41469 (2012).
35. Liu, X. *et al.* Serum Antibody Repertoire Profiling Using In Silico Antigen Screen. *PLoS One* **8,** e67181 (2013).
36. Burks, A. W. *et al.* Mapping and mutational analysis of the IgE-binding epitopes on Ara h 1, a legume vicilin protein and a major allergen in peanut hypersensitivity. *Eur. J. Biochem.* **245,** 334–339 (1997).
37. Shreffler, W. G., Beyer, K., Chu, T. H. T., Burks, A. W. & Sampson, H. A. Microarray immunoassay: Association of clinical history, *in vitro* IgE function, and heterogeneity of allergenic peanut epitopes. *J. Allergy Clin. Immunol.* **113,** 776–782 (2004).
38. Burks, A. W. *et al.* Identification of a major peanut allergen, Ara h I, in patients with atopic dermatitis and positive peanut challenges. *J. Allergy Clin. Immunol.* **88,** 172–179 (1991).
39. Gould, H. J. *et al.* The biology of IGE and the basis of allergic disease. *Annu. Rev. Immunol.* **21,** 579–628 (2003).
40. Hamilton, R. G., MacGlashan, D. W. & Saini, S. S. IgE antibody-specific activity in human allergic disease. *Immunol. Res.* **47,** 273–284 (2010).
41. Scott, B. M. *et al.* Phage display of the serpin alpha-1 proteinase inhibitor randomized at consecutive residues in the reactive centre loop and biopanned with or without thrombin. *PLoS One* **9,** e84491 (2014).
42. Bratkovič, T. Progress in phage display: Evolution of the technique and its applications. *Cell. Mol. Life Sci.* **67,** 749–767 (2010).
43. Smith, G. P. & Petrenko, V. A. Phage Display. *Chem. Rev.* **97,** 391–410 (1997).
44. Kim, T. *et al.* MUSI: An integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic Acids Res.* 40, e47 (2012).
45. Lin, J. *et al.* A bioinformatics approach to identify patients with symptomatic peanut allergy using peptide microarraundery immunoassay. *J. Allergy Clin. Immunol.* **129,** 1321–1328 e5 (2012).
46. Bøgh, K. L. *et al.* IgE versus IgG4 epitopes of the peanut allergen Ara h 1 in patients with severe allergy. *Mol. Immunol.* **58,** 169–176 (2013).
47. Simons, F. E. R. *et al.* World Allergy Organization anaphylaxis guidelines: summary. *J. Allergy Clin. Immunol.* **127,** 587–93, e1–22 (2011).
48. Bøgh, K. L., Barkholt, V., Rigby, N. M., Mills, E. N. C. & Madsen, C. B. Digested Ara h 1 loses sensitizing capacity when separated into fractions. *J. Agric. Food Chem.* **60,** 2934–2942 (2012).
49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
50. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets. *Protein Sci.* **1,** 409–417 (1992).
51. Thomsen, M. C. F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.* 40, W281–W287 (2012).

Acknowledgements

The authors thank Neslihan Bicen and the staff at the DTU Multi-Assay Core (DMAC) facility (DTU, Denmark) for useful technical assistance regarding high-throughput sequencing. This work was supported by The Danish National Advanced Technology Foundation (grant number 128-2012-1), Dr. Techn. A.N. Neergaard and Wife's Foundation and the Technical University of Denmark.

Author Contributions

A.C., K.L.B. and M.D. designed the study. A.C. performed experiments. J.V.K., C.S.H., M.N. and O.L. designed the bioinformatics pipeline. E.S., J.P., C.S.H. and F.D.M. designed and performed peptide micro-array validations. A.C., J.V.K., C.S.H. and M.D. analysed the data. N.M.R., T.E. and Z.S. provided and characterised crucial reagents. A.C., J.V.K. and C.S.H. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: Eric Sullivan and Jigar Patel are full-time employees of Roche-NimbleGen. The other authors declare no competing financial interests.

How to cite this article: Christiansen, A. *et al.* High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5**, 12913; doi: 10.1038/srep12913 (2015).

Chapter 2

High-throughput sequencing enhanced phage display for discovery of peptidic anti-cobratoxins

Introduction

Snakebite envenoming is considered one of the most neglected tropical diseases in the world[1,2]. Even though the incidence is highest in rural areas of Asia and Sub-Saharan Africa, snakebites represent a global problem[3]. It is estimated that between 0.5 and 2 million people suffer from venomous snakebites every year, resulting in 20,000-100,000 deaths[3,4].

Cobras (Naja genus) are members of the elapid family and are known for their neurotoxic venom. Among these, the Monocled cobra, *Naja kaouthia*, is widespread in Southern Asia and is responsible for a substantial part of the recorded bites[5,6]. The clinical manifestations of an envenoming by *Naja kaouthia* is characterized by neuromuscular paralysis, which can result in respiratory failure, and by local tissue damage, including swelling, blistering, and necrosis[7]. The most abundant and toxic component in *Naja kaouthia* venom is the α-neurotoxin protein, α-cobratoxin (molecular weight: 7.8 kDa) [6,8,9]. α-cobratoxin inhibits the nicotinic acetylcholine receptor (nAChR) at the endplate of muscle fibers leading to paralysis (see Figure 1).

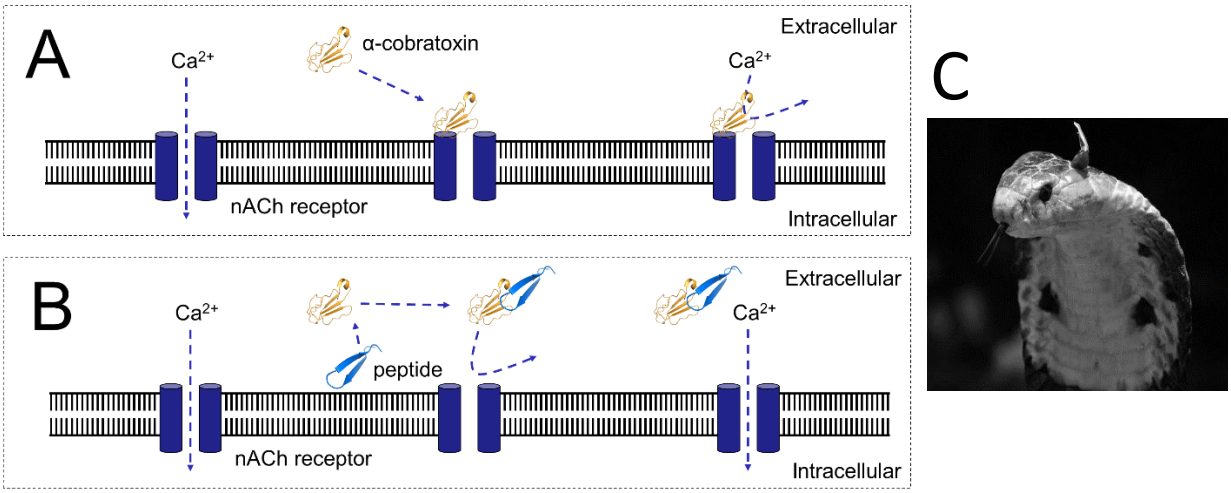


Figure 1. The principle behind the physiological response to α-cobratoxin. **A)** α-cobratoxin inhibits the nicotinic acetylcholine receptor (nAChR) at the endplate of muscle fibers leading to paralysis. **B)** Supposed action of a peptide-based inhibitor. The inhibitor binds α-cobratoxin and prevent the toxin from inhibiting the nAChR; as a result, the muscle functionality is maintained. **C)** A photograph of a *Naja kaouthia* (Monocled cobra)[9]. Source: Figure A and B have been provided by Andreas Laustsen.

The only effective treatment to snakebite envenoming involves the administration of antivenom[2,10]. However, antivenoms are produced by animal immunization and are therefore associated with a high

risk of adverse effects[11]. Even though the numbers of snakebite incidences seem high, there are only limited commercial interests and research efforts going into antivenoms[1,12]. There has been an increasing interest in using antibody fragments or smaller proteins and peptides for treatment. Among the proposed benefits of synthetic peptides are fewer adverse effects, lower cost, a more homogenous product as well as an improved distribution profile[13].

Study objective

The overall objective of the present study was to discover novel peptide-based antitoxins that can be used for development of a synthetic antivenom against neurotoxins from cobras. The focus was on α-cobratoxin since it is the most abundant and toxic component in monocled cobra venom[6,8,9]. Other cobra species have similar neurotoxins and an α-cobratoxin inhibitor may also neutralize their toxins. Peptidic drug candidates were discovered by phage display. The strategy was to identify peptides that bound the toxin by biopanning phages displaying random peptide libraries against α-cobratoxin. Next, the peptide binders were functionally tested to determine whether they inhibited the toxin.

From a methodological perspective, there was an additional aim of the study. After we had demonstrated that phage display coupled with HTS enabled epitope mapping of patient sera (Chapter 1), we sought to investigate whether HTS could enhance a different application of phage display, specifically the identification of peptide binders. The hypothesis was that a deeper sequencing of the phage library could help identity peptides that were missed by traditional phage display. In addition, this study also provided the opportunity to compare the output of traditional phage display screening with that of HTS-coupled phage display screening. Since the study was not designed for HTS analysis, there were some inherent limitations. First of all, selected phages were not available for all selection rounds and, secondly, the phages used as controls were derived from samples that were selected against toxins. Consequently, α-cobratoxin binding peptides would be expected in the control samples, making it impossible to use the controls for direct contrasting, similar to the approach in Chapter 1. Nonetheless, it was assessed that the experiments could provide new insights into the potential advantages of deep sequencing selected phage pools.

Methods

Phage display screening

Each well in an amine-binding plate 'ReactiBind' (Thermo Scientific) was coated with 100 µL 10.0 µM α-Cobratoxin (99%, Latoxan) in PBS. Control wells were coated with 100 µL PBS. Amine-binding wells were capped by adding 200 µL 10% (v/v) ethanolamine. The incubation time was 1 h at room temperature (RT) followed by incubation over night at 4 °C. Subsequently, the wells were washed 3x with 200 µL PBS w/ 0.05% Tween® 20 and 2x with 200 µL PBS. The wells were then blocked by addition of 100 µL Blocking Buffer (0.1 M $NaHCO_3$ pH 8.6, 0.02% NaN_3, 0.1 µg/mL streptavidin (streptavidin addition was unintended, since the selection process did not involve biotinylation), 5 mg/mL Bovine Serum Albumin (BSA)) and incubated at RT on a plate shaker for 2 h. MaxiSorp plates were not coated, but they were blocked in a similar way. Control wells were blocked and coated in the same manner as the test wells, except ReactiBind plates were coating with PBS instead of α-Cobratoxin, as described above.

For the first biopanning round, a stock solution of the libraries TriCo-16™ Phage Display Peptide Library and TriCo-20™ Phage Display Peptide Library (Creative Biolabs) was used. 2.5 µL of each library was mixed with 105 µL Blocking Buffer and incubated for 30 min at RT. To remove plastic binding phages, 100 µL of this phage solution was transferred to a blocked well in a MaxiSorp plate (Thermo Scientific) and incubated for 1 h at RT. After blocking of the ReactiBind wells, these were washed 10x with 200 µL PBS w/ 0.05% Tween® 20 and 2x with PBS to remove excess blocking agent. 100 µL of the phages from the MaxiSorp plate were transferred to the ReactiBind wells and incubated for 1 h at RT. To remove unbound phages wells were washed 10x with PBS w/ 0.05% Tween® 20 and 2x with PBS. The bound phages were eluted by incubating for 10 min with 200 µL 100 mM triethylamine followed by neutralization with 400 µL 1 M Tris-HCl buffer, pH 7.4. Between each selection round the eluted phages were amplified in 8-10 mL 2x YT media (Sigma) at 37 °C overnight by infecting *E. coli* TG1. Amplified phages were precipitated using PEG/NaCl and titered. Selection round 2-5 were carried out in a similar manner using the precipitated phages from the previous round. Specifically, 115 µL out of a mixture of 125 µL phage and 125 µL Blocking Buffer was added to a blocked well in a MaxiSorp plate and incubated for 1 h at RT and then subsequently used in the ReactiBind plates. Note that increasing phage titer values were used to verify a successful enrichment in each selection round compared to the previous round, however, the specific titer values were not used to extract an amount of phages for the following selection round; instead a constant volume was used. In parallel to the biopanning selection against α-

cobratoxin, control reactions were also carried out. For each control biopanning the eluate from the previous phage selection round were selected against PBS coated wells and handled as described above. Selection of the plate type and blocking procedures were the result of optimization attempts by the collaborator, after initial attempts using only MaxiSorp plates were unsuccessful.

Isolation of single-clones and ELISA

After titering of selected phage eluates, 24 single plaques were transferred to Eppendorf tubes and amplified over night. After 10 min at 8,000 x g centrifugation, 125 µL of the supernatant was used for ELISA. For the ELISA, wells were coated with α-cobratoxin or PBS. The following day, the wells were emptied and washed, blocked and washed again. Phage eluate from each clone was mixed 1:1 with Blocking Buffer. 100 µL of the phage eluate in Blocking Buffer was added to α-cobratoxin and PBS coated wells. The plate was incubated for 1 h at RT on a plate shaker. After incubation, the wells were washed with 10x 100 µL PBS w/ 0.05% Tween® 20 and 2x PBS to remove unbound phages. 100 µL of a solution of a 1‰ monoclonal Horseradish Peroxidase (HRP) conjugated Anti-M13 phage antibody (GE Healthcare) in Blocking Buffer was added to each well and incubated on a plate shaker for 1 h and washed subsequently. One 2 mg *ortho*-Phenylenediamine (OPD) tablet (DAKO) was dissolved in 3 mL H_2O and 5 µL H_2O_2 as a substrate for HRP and applied as instructed by the manufacturer. 100 µL OPD solution was added to each well and the plate was left for up to 30 min or until sufficiently strong color change was observed. Addition of 100 µL 0.5 M H_2SO_4 to all the wells at the same time stopped the reaction and absorbance at 490 nm was measured (VersaMax Tunable Microplate reader, Molecular Devices). Empty wells were used as blank controls for substraction of signal.

The 12 clones that gave the highest responses in the ELISA were selected for further processing. They were amplified overnight and PEG/NaCl precipitated. The ssDNA was isolated and Sanger sequenced at Eurofins Genomics (Eurofins MWG Operon).

High-throughput Sequencing

Amplified phage eluates from each available selection round were subjected to PCR using the Phusion High-Fidelity DNA polymerase (New England Biolabs). The samples were extracted from the same tube that was used above (i.e. the eluates that gave rise to clones that were Sanger sequenced, were used as

template for PCR reactions). Samples selected against toxin as well as control samples were handled in a similar manner. To mitigate PCR bias, the PCR reactions were carried out in triplicates and the number of cycles was limited to 25. The PCR was carried out as recommended by the manufacturer, using 25 µL reactions with 1 µL 1:10 diluted phage eluates as template and an annealing temperature of 55 °C. The variable part of the phage library was specifically amplified using 8 barcoded primers. The corresponding PCR fragment was extracted from a 2% agarose gel using the Qiaquick Gel Extraction Kit (Qiagen). The triplicate PCR reactions were pooled prior to the gel extractions. The DNA concentration was measured on a Qubit fluorometer (Life) and amplicon integrity was confirmed using a 2100 Bioanalyzer (Agilent). Next, sequencing adaptors were ligated onto the DNA fragments using the Ion Plus fragment library kit (Life), and the fragment size and concentration was again validated on a Bioanalyzer (Agilent) and Qubit fluorometer (Life), respectively. Finally, the DNA was ligated to Ion Sphere Particles (ISPs) and subjected to emulsion PCR according to the Ion OneTouch 200 protocol. The ISPs were loaded onto an Ion 318 chip and sequenced on an Ion Torrent PGM, as described by the manufacturer (Ion Torrent sequencing kit 200 v2).

Converting sequencing reads to peptides

DNA reads obtained from the HTS was first filtered by sample barcodes. Reads that did not match any of the barcodes, both forward and reverse complement, were removed. The DNA reads were then translated in all six reading frames and aligned (using BLAST) to the protein regions of the phage protein that flank the displayed 16-mer or 20 peptide. Then 16-mer and 20-mer peptides and their corresponding DNA sequences were obtained from reads where the protein sequence aligned to both flanking regions with exactly a 16- or 20-mer gap. The peptides were further processed by removing peptides where at least one DNA base pair (bp) coding for the peptide were assigned a Phred quality score below 15.

Two-electrode voltage clamp experiments

Rat α7 nAChR (containing the L247V mutation to limit desensitization) and human RIC-3 (resistant to inhibitor of cholinesterase; enhances α7 nAChR expression) plasmid cDNAs were linearized with NotI or NheI (New England Biolabs), and cRNA was synthesized with the SP6 or T7 mMessage Machine

transcription kits (Life Technologies). After surgical removal from Xenopus laevis frogs, oocytes were separated and defoliculated in 0.1 mg/ml Type II collagenase and then stored at 18 °C in OR2 medium (82.5 mM NaCl, 2.5 mM KCl, 1 mM $MgCl_2$, 5 mM HEPES/NaOH pH 7.4). Stage V or VI oocytes were injected with 4 ng α7 nAChR cRNA and 2 ng RIC-3 cRNA and incubated at 18 °C for 2 days in OR2 medium.

For the electrophysiological analysis, oocytes were continuously perfused with bath solution (96 mM NaCl, 2 mM KCl, 1.8 mM $CaCl_2$, 1 mM $MgCl_2$ 5 mM HEPES/NaOH pH 7.4) and two electrode voltage clamped (Warner OC-725C amplifier, Molecular Devices 1550 digitizer). Currents were measured in response to 100 µM acetylcholine alone and then three times in the presence of 40 nM α-cobratoxin or 40 nM α-cobratoxin and 100 µM peptide. This was followed by two to three additions of acetylcholine alone ("wash"). Peptide and α-cobratoxin were mixed and incubated at RT for 30 minutes before experiments. For each experiment (i.e. individual oocytes), all measurements were normalized to the response to acetylcholine alone.

Results

Selection of α-cobratoxin binders by phage display screening

Phages from 16- and 20-mer random peptide libraries were selected for binding to α-cobratoxin over 5 consecutive selection rounds. The eluted phages were titered after all selection rounds except for the first. For all panning rounds the titer for the selected sample was higher than a parallel control selection. The specific ratio of binder/control titers were 24, 203, 755 and 175 for selection round 2 through 5. The increasing ratios between each selection round (except for the 5^{th} round) indicated that a specific α-cobratoxin binder had been amplified.

Next, 24 individual phage clones were isolated and tested by ELISA. As shown in Figure 2, the response varied greatly among clones. The negative control gave a very low signal. The 12 clones with the highest absorbance were selected for further processing. Sanger sequencing of the clones revealed that all of them expressed 16-mer peptides. Of the 12 clones, 11 expressed the same peptide, PEP1 (sequences have been anonymized due to patent considerations). The remaining clone expressed a highly similar peptide, PEP2, which had 75% sequence identity to PEP1 (12 out of 16 amino acids were identical).

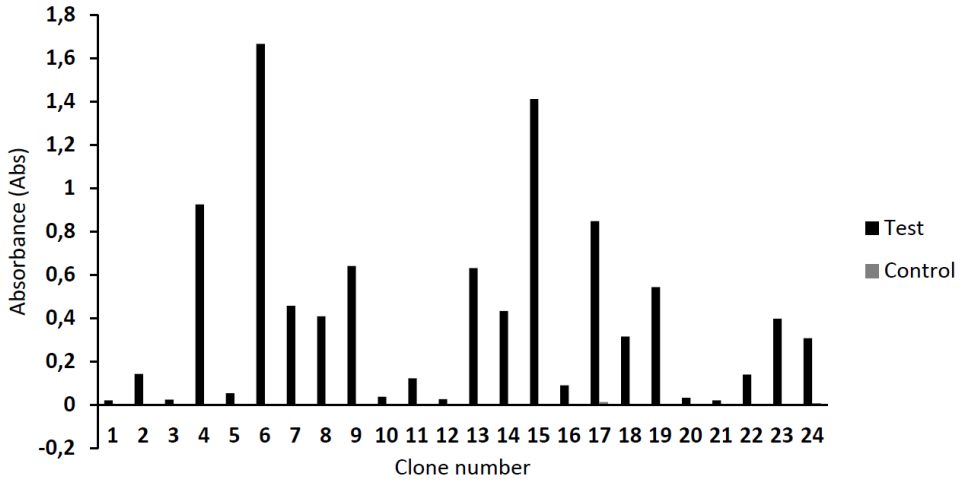


Figure 2. ELISA on isolated phage clones selected as α-cobratoxin binders. The plates were amine-binding plates and the phages were detected using HRP-labeled Anti-M13 antibodies. The absorbance measured at 490 nm is shown after the subtraction of the mean absorbance of five empty wells. The control wells were coated with PBS.

High-throughput sequencing of the selected phage libraries

Amplified phage eluates from biopanning round 1, 2, 3 and 5 were available for HTS along with round 2, 3, 4 and 5 for the control selections. The libraries were PCR amplified with barcoded primers, pooled and sequenced using an IonTorrent PGM. The phage sequences were extracted, split according to barcodes and converted to peptide reads. For unknown reasons, the sequencing failed for the 5th biopanning round against α-cobratoxin as well as the 2nd biopanning round for the control. The number of peptides allocated to these samples were <400 compared to >100,000 for the remaining six samples. Accordingly, these samples were disregarded in further analyses. The 10 most frequent peptides identified in the remaining samples are presented in Table 1. A total of 1,469,811 sequences were obtained representing an average of 224,969 per sample (range 110,503-499,660). The overall quality of the sequencing run was relatively low: if the quality (Phred) score cutoff was increased from 15 to the previously used 20, the approved number of sequences dropped to 674,989. Regarding the phage selection process, the number of unique peptides dropped with each subsequent round (see Table 1). Specifically, from 4938 to 410 for round 1 to 3 for the toxin-selected samples and similarly from 413 to 108 from round 3 to 5 in the control samples. Sequences that are designated PEP are peptides that were selected for further analyses, see below.

Selection of peptides for further analyses

The prominent clone identified by traditional phage display, PEP1, was also quite prevalent in the sequenced libraries. It was not present in the sequences from selection round 1 and 2, but it was the most frequent peptide in the 3rd selection round. In the control selections, PEP1 was very prevalent in round 4 and 5, ranking second and first in frequency, respectively. It could not be detected in the 3rd control selection round. In contrast, the similar peptide, PEP2, which was identified in a single clone in the conventional phage display, could not be identified in the deep sequenced libraries.

That a known peptide binder, PEP1, was the most frequent peptide in the 3rd selection round suggested that frequent peptides in this round could be potential binders. Accordingly, three additional peptides (PEP3-5) ranking 2-4 in frequency were selected for further studies. Another peptide, PEP6, was included since it ranked 7 and was consistently prevalent in the later control selections. The most frequent peptide in the first selection round, PEP7, was also selected. Finally, a peptide ranking 6 in the 3rd toxin selection round, PEP8, was included since it was highly similar to PEP4 differing only in 3 C-terminal amino acids out of the 20-mer peptide. Inspired by the findings in Chapter 1, the data was investigated for putative peptide motifs. However, no major motifs could be identified (data not shown).

Toxin Round 1		**Toxin Round 2**		**Toxin Round 3**	
PEP7 (20)	8.60%	PEP4 (20)	94.89%	PEP1 (16)	49.99%
PEP4 (20)	5.63%	PEP5 (20)	2.52%	PEP4 (20)	45.14%
YPGSPTQYPSSMHEYHSSSE	2.65%	PEP3 (20)	1.68%	PEP3 (20)	3.83%
SKFHFESYFDMVPYRQAEAN	1.86%	GSWERVIHPAQRVAWHPYQV	0.04%	PEP5 (20)	0.78%
GSWERVIHPAQRVAWHPYQV	1.64%	NVVFKGPHSNPPVWLS	0.04%	HVYSVERMTDVSNQIWIDTR	0.02%
SSFMLEEPWQPQHRFRLSMK	1.60%	SPMHLVATPARHFPAT	0.03%	PEP8 (20)	0.01%
NVVFKGPHSNPPVWLS	1.45%	SSFMLEEPWQPQHRFRLSMK	0.02%	PEP6 (16)	0.01%
TDRHVHMPPALGQYTI	0.99%	SYRIADDRPSDHIMYAPTMN	0.02%	SPMHLVATPARHFPAT	0.01%
QPHKFPEGWHATTKLTLKVR	0.75%	SKFHFESYFDMVPYRQAEAN	0.02%	GSWERVIHPAQRVAWHPYQV	0.01%
NWSTHLYANKSHAPYSQPLF	0.67%	QPHKFPEGWHATTKLTLKVR	0.01%	SSFMLEEPWQPQHRFRLSMK	0.01%
Unique peptide #:	4938	Unique peptide #:	621	Unique peptide #:	410
Total sequence #:	110503	Total sequence #:	192216	Total sequence #:	499660

Table 1. The ten most frequent peptides in selection round 1-3 against α-cobratoxin. The number of unique peptides as well as the total number of sequences are included. (16) or (20) designates a 16- or 20-mer peptide.

Control Round 3		Control Round 4		Control Round 5	
PEP4 (20)	95.43%	PEP4 (20)	81.09%	PEP1 (16)	47.55%
PEP6 (16)	1.48%	PEP1 (16)	11.46%	PEP4 (20)	46.43%
PEP3 (20)	0.64%	PEP3 (20)	3.44%	PEP3 (20)	3.61%
PEP5 (20)	0.39%	PEP6 (16)	2.22%	PEP6 (16)	1.88%
SPMHLVATPARHFPAT	0.23%	PEP5 (20)	0.53%	PEP5 (20)	0.16%
GSWERVIHPAQRVAWHPYQV	0.12%	SPMHLVATPARHFPAT	0.31%	GHPVKNTQRYHHAIGE	0.07%
SSFMLEEPWQPQHRFRLSMK	0.09%	GHPVKNTQRYHHAIGE	0.15%	GVHQYSRPTVPSYLWTSGQH	0.05%
NVVFKGPHSNPPVWLS	0.09%	DGYKLQTSLDWQMWNP	0.11%	SSFMLEEPWQPQHRFRLSMK	0.03%
HSKWLSEWTHGWEQFY	0.07%	HSIARGPEFGHYTDRL	0.08%	KPPNAPLMLWQSDFVHKDVA	0.03%
EHYWEWMHFNTHVYTP	0.06%	HHWQAPQWDNMSMWRT	0.07%	NKHWPVYETWQFSDYH	0.02%
Unique peptide #:	413	Unique peptide #:	136	Unique peptide #:	108
Total sequence #:	342131	Total sequence #:	157212	Total sequence #:	168089

Table 2. The ten most frequent peptides in control selection round 3-5 against PBS. Control selections were performed in parallel with regular toxin selection by using the eluate from the previous toxin selection round and panning it against PBS-coated wells. The number of unique peptides as well as

Functional testing of selected peptides

The selected peptides were all tested for binding to α-cobratoxin by isothermal titration calorimetry (ITC). For the originally identified PEP1 along with PEP3 and PEP4 binding was confirmed (data not shown). Even though PEP8 exhibited high similarity to PEP4, binding could not be confirmed for this peptide. All peptides were tested in two electrode voltage clamp experiments (TEVCs), see Figure 3 & 4. The α-cobratoxin inhibited the nAChR in *Xenopus laevis* oocytes leading to a decrease in measured currents. As shown in Figure 3, PEP1 and PEP3 prevent the inhibition caused by α-cobratoxin, whereas PEP4 enhances both the onset and wash-out of inhibition. The latter may suggest that PEP4 is a modulator rather than an inhibitor. The predicted structures for PEP1, 3 and 4 are shown in Figure 3.

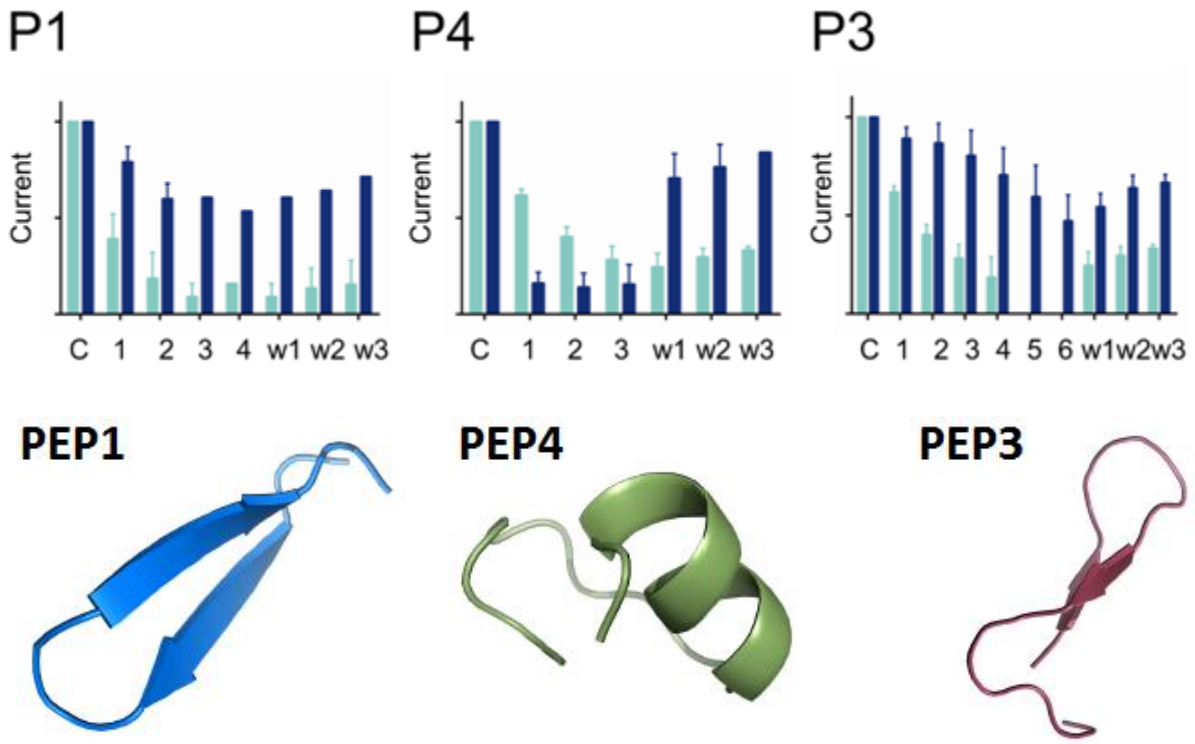


Figure 3. Top half: Peptides prevent α-cobratoxin from inhibiting nAChRs in *Xenopus laevis* oocytes in voltage clamp experiments (TEVC). 100 µM acetylcholine-gated currents were recorded alone (control, "C"); in the continued presence of either 40 nM α-cobratoxin alone (light blue bars, "1-3") or 40 nM α-cobratoxin and 100 µM peptide (dark blue bars, "1-3"); and then alone again (wash, "w1-w3"). The observations are described in the main text. The normalized mean from 3-5 experiments is shown. Error bars denote standard error of the mean. Bottom half: The peptide structures as predicted by PEP-FOLD (http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD/). Figures were supplied by Andreas Laustsen and modified.

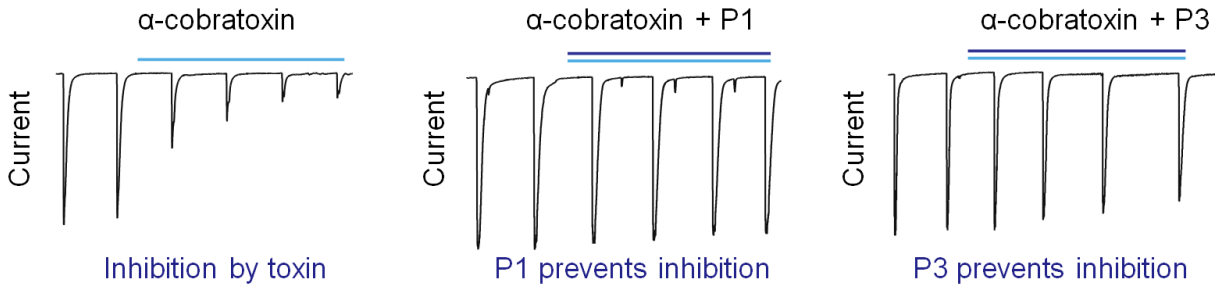


Figure 4. Measured ion currents through the nAChR in sample *Xenopus laevis* oocytes in a TEVC assay. The current in an assay with only α-cobratoxin is shown to the left and the inhibition of ion current flow caused by addition of PEP1 (P1) and PEP3 (P3) is shown in the middle or to the right, respectively. Figures were supplied by Andreas Laustsen.

Discussion

In this study, phage libraries displaying random peptides were selected for binding to the α-cobratoxin. Traditional clone isolation and testing identified two peptides, one of them (PEP1) was confirmed as a toxin binder and inhibitor. Deep sequencing of the selected phage eluates revealed a much greater diversity of phages. Six frequent peptides were selected for further testing, and two of them (PEP3 and PEP4) were confirmed as either inhibiting or modulating the effect of α-cobratoxin.

HTS analysis of the selected libraries provided insights into the phage selection process. In some ways the development was similar to what was reported previously in Chapter 1. The number of unique peptides identified in each selection round dropped with each subsequent round, and were generally comparable to what we had previously observed[14]. As before, the greatest drop in diversity was observed from round 1 to 2. The number of unique peptides kept falling in round 4 and 5 in the control setting, however the drop from round 4 to 5 was relatively small. These observations fit well with the general affinity selection concept in phage display where it is expected to observe an increase in specificity but a decrease in diversity with each selection round[15,16].

Examining the changes in frequency proportions, i.e. the frequencies of the most prevalent peptides, there were some differences from the study in Chapter 1. In the previous study, the frequency of the most prevalent peptide increased with each subsequent round and ended up dominating each sample. In this study, there were again dominant sequences but the development was more heterogeneous. E.g., the most prevalent peptide in the 2nd selection round had a frequency of about 95%, whereas the 3rd round the two most frequent peptides had frequencies of about 45% and 40%. Along these lines, there appeared to be slightly less stability in the prevalent peptides (e.g. top 10) compared to the earlier study: Three peptides remained in the top 10 from selection round 1 to 3. This was the case for an average of six peptides in the previous study[14].

After the successful identification of α-cobratoxin binders (PEP1, 3 and 4), it is relevant to re-examine their prevalence throughout the selection procedure. Overall, prevalence seemed to be a good indication for α-cobratoxin binding. In selection round 1, 2 and 3 there were 1, 2 and 3 binders among the three most frequent peptides, respectively. The peptide identified by traditional phage display, PEP1, was the last to emerge. These observations were mirrored by the results for the control selections, though there was a delay in the emergence of binders. This observation was to be expected, since the control selections were carried out using phages from the previous α-cobratoxin selection

round that were subjected to a single unspecific selection round. The unspecific selection likely decreased the frequency of the α-cobratoxin binders. For 4 out of the 6 peptides selected after HTS analysis, neither a binding nor a modulating effect could be confirmed. Most of the non-binding peptides were the least frequent of the examined peptides in the 3rd selection round, possibly supporting that prevalence was an appropriate indicator for α-cobratoxin binding.

PEP2, which was identified in traditional phage display and was similar to PEP1, could not be confirmed as a binder despite the high degree of sequence similarity. This was also the case for PEP8, which was selected for testing based on its similarity with PEP4. The similarity of PEP2 and 8 to known binders may suggest that they were binders in the phage experiments. However, when the peptides were separated from their phage context their structures might have altered, resulting in diminished binding capability. It is also worth noting, that the two peptides (PEP1 and PEP2) identified by traditional phage display were 16-mers while HTS analysis revealed a mixed phage population expressing both 16- and 20-mers.

Though HTS enabled the identification of additional binders, there were some limitations to the study. This was mainly a result of the fact that the study had not been designed with HTS in mind. First of all, samples were unavailable for some selection rounds. This issue expanded with the unsuccessful sequencing of two available samples. The missing samples restricted the downstream analysis and made it more difficult to evaluate the selection process comprehensively. Another result of the experimental design, was that control samples were derived from samples that were selected against toxins. Consequently, α-cobratoxin binding peptides could be identified in the control samples, and the samples could not be used for direct contrasting similarly to the process performed in Chapter 1.

Overall, the HTS analysis of the selected phage libraries proved to be a valuable approach since it enabled the identification of α-cobratoxin binders that were not identified in the original phage display experiments. The study thus confirmed the initial hypotheses. The study demonstrated the ease of transferring HTS enhanced phage display beyond epitope mapping to another major application of phage display, namely the identification of binding peptides. In this study, peptide frequency appeared to be a relevant selection parameter, which indicates that the applied selection pressure worked as intended. Further studies of the peptide inhibitors include truncation and mutational experiments to identify the binding core of the peptides. Finally, the capability of the inhibitors may be tested *in vivo* in rodent models. If sufficiently good inhibitors can be obtained, such peptide-based anti-cobratoxins could serve as crucial components to neutralize the neurotoxic effects of cobra envenomings.

Contributions

Andreas Laustsen and Brian Lohse from the Department of Drug Design and Pharmacology, University of Copenhagen, participated in the design of the study. Andreas Laustsen managed the study. Jónas Johanneson performed the phage display experiments and the ELISAs, and Timothy Lynagh contributed with the TEVC data, both from the Department of Drug Design and Pharmacology, University of Copenhagen. The DTU Multi-Assay Core facility performed the IonTorrent sequencing. Anders Christiansen designed the HTS part of the study, designed primers, performed the sample preparation for deep sequencing and analyzed the HTS data.

References

1. Brown, N. I. Consequences of neglect: analysis of the sub-Saharan African snake antivenom market and the global context. *PLoS Negl. Trop. Dis.* **6,** e1670 (2012).

2. Gutiérrez, J. M., León, G. & Burnouf, T. Antivenoms for the treatment of snakebite envenomings: the road ahead. *Biologicals* **39,** 129–42 (2011).

3. Kasturiratne, A. *et al.* The global burden of snakebite: a literature analysis and modelling based on regional estimates of envenoming and deaths. *PLoS Med.* **5,** e218 (2008).

4. Chippaux, J. P. Snake-bites: appraisal of the global situation. *Bull. World Health Organ.* **76,** 515–24 (1998).

5. Viravan, C., Veeravat, U., Warrell, M. J., Theakston, R. D. & Warrell, D. A. ELISA confirmation of acute and past envenoming by the monocellate Thai cobra (Naja kaouthia). *Am. J. Trop. Med. Hyg.* **35,** 173–81 (1986).

6. Kulkeaw, K. *et al.* Human monoclonal ScFv neutralize lethal Thai cobra, Naja kaouthia, neurotoxin. *J. Proteomics* **72,** 270–82 (2009).

7. Wongtongkam, N., Wilde, H., Sitthi-Amorn, C. & Ratanabanangkoon, K. A study of Thai cobra (Naja kaouthia) bites in Thailand. *Mil. Med.* **170,** 336–41 (2005).

8. Richard, G. *et al.* In vivo neutralization of α-cobratoxin with high-affinity llama single-domain antibodies (VHHs) and a VHH-Fc antibody. *PLoS One* **8,** e69495 (2013).

9. Laustsen, A. H. *et al.* Snake venomics of monocled cobra (Naja kaouthia) and investigation of human IgG response against venom toxins. *Toxicon* **99,** 23–35 (2015).

10. Lalloo, D. G. & Theakston, R. D. G. Snake antivenoms. *J. Toxicol. Clin. Toxicol.* **41,** 277–90; 317–27 (2003).

11. León, G. *et al.* Pathogenic mechanisms underlying adverse reactions induced by intravenous administration of snake antivenoms. *Toxicon* **76,** 63–76 (2013).

12. Brown, N. & Landon, J. Antivenom: the most cost-effective treatment in the world? *Toxicon* **55,** 1405–7 (2010).

13. Vlieghe, P., Lisowski, V., Martinez, J. & Khrestchatisky, M. Synthetic therapeutic peptides: science and market. *Drug Discov. Today* **15,** 40–56 (2010).

14. Christiansen, A. *et al.* High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5,** 12913 (2015).

15. Smith, G. P. & Petrenko, V. A. Phage Display. *Chem. Rev.* **97,** 391–410 (1997).

16. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16,** 1776–1803 (2011).

Chapter 3

Towards epitope mapping of therapeutic candidate antibodies in immunized animal sera

Introduction

Antibodies are becoming widely used in disease therapy, e.g. in the treatment of cancer or autoimmune diseases[1–3]. New antibody-based therapeutics are approved on a regular basis and the annual revenue is estimated to be at least 20 billion dollars[1,3]. Identification and development of therapeutic antibodies is a long process. One way to identify antibody candidates (outlined in Figure 1) is to immunize animals (possibly chimeric) with a therapeutic target, isolate B-cells and screen the B cells for relevant antibody production[2]. When suitable antibodies have been identified, the candidates undergo extensive development (including humanization if the antibody library is of animal origin) and testing before they are suited for clinical trials[1]. Epitope mapping is often a part of the antibody characterization process since it facilitates antibody engineering and assists in protecting intellectual property rights[1,4]. Most importantly, it aids the selection of lead antibody candidates e.g. facilitating the selection of antibodies that target a diverse set of epitopes or antibodies that target or avoid specific known epitopes (e.g. well-known epitopes from published antibodies or epitopes that interact with a specific ligand).

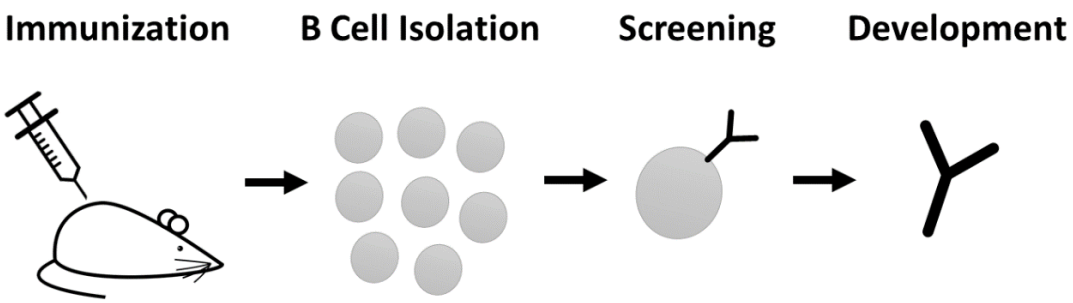


Figure 1. A schematic overview of a therapeutic antibody development process. First, an animal, e.g. a mouse, is immunized with the therapeutic target. B cells are then isolated by cell sorting samples from the immunized animal. The selected B cells are then screened for antibodies that bind the target. The identified antibody candidates are subjected to further development and testing. Source: This figure incorporates elements under the Creative Commons License.

Study Objective

In this study, the overall objective was to identify the epitopes of a therapeutic antibody candidate (the identity of the target in question is not disclosed due to confidentiality agreements). This was a pilot study aiming at analyzing sera from immunized animals and deducing the epitope patterns of the target. Concurrent screening of extracted B cells were to assist identification of true antibody candidates.

Consequently, it would be known whether a given animal produced antibodies of interest. Accordingly, one goal was to examine whether there were characteristic epitope patterns, e.g. the number and diversity of epitopes, which characterized animals that produced useful antibody candidates. After a successful pilot study, samples from additional time points would be analyzed to characterize the temporal development of the epitope patterns after immunization. Finally, the B cells would also be characterized by HTS, thus providing knowledge on the antibody sequence diversity. In the end, the epitope patterns could be compared to the antibody sequence output.

From a methodological perspective, having demonstrated that phage display coupled with HTS could be used to map epitopes in sera from allergic patients (Chapter 1), we wanted to confirm the approach on a different set of samples. The animal samples were markedly different from the human sera: The hypothesis was that the relevant antibody levels would be much higher in immunized animals versus the low level of IgE observed in allergic patients. This could, in principle, simplify the analysis and allow the detection of several epitopes in a single sample. Finally, the previously applied IonTorrent sequencing platform had exhibited some limitations with regards to data quality, and accordingly we also sought to investigate advantages of transferring to the Illumina MiSeq sequencing platform.

It should be noted that due to confidentiality agreements the animal platform (as well as the antibody isotype) is not described in this research note.

Methods

Animal samples

Six animals were immunized with the target. ELISA on samples at Day 19 confirmed that all animals produced antibodies that bound the target. Samples from two of the animals with highest responses (highest reactivity on ELISA), Animal 1 and 6, as well as an animal in the low-response group, Animal 2, were selected for this study. The reason to select high and low responders was to determine whether any correlation could be observed between the ELISA response and epitope patterns. Serum was available from before immunization, Day 0, which would act as a baseline control since the animal did not have antibodies against the target. Another serum sample was obtained 19 days after the initial immunization. Details have been omitted from this section due to confidentiality agreements.

Phage display - Selection of target mimicking peptides

The general approach is similar to the phage display experiments described in Chapter 1 and the principle is illustrated in Figure 1 in that chapter. In brief, 6 × 200 µL (6 mg) of M-280 tosylactivated Dynabeads (Invitrogen) were coated with 20 µg of antibodies capturing the most frequent antibody isotype in the animal for 48 h at 37 °C, and blocked with 0.5% SMP in PBS. 250 µL serum was diluted in PBS with 0.5% SMP and 0.05% Tween 20, and incubated over night at 4 °C with 200 µL of the coated bead suspension. After extensive washing, 50 µL of the coated beads were used immediately for the first round of selection, while the remaining 150 µL were stored for later selection rounds.

In the first selection round, 50 µL of the beads coated with antibodies were blocked in 1 mL of 2% SMP in PBS for 1 h at room temperature and washedcomprehensively . The beads were then incubated overnight at 4 °C with 10 µL (≈ 2×10^{11}) phages from the Ph.D.-7™ library (phages displaying random heptamer peptides, source: New England BioLabs). This was followed by washing and negative selection as follows: 1 h incubation in 10 mL PBS with 0.5% SMP and 0.05% Tween 20 and 8 washing steps each consisting of 10 min incubation in 10 mL PBS with 0.05% Tween 20. Next, phages of interest were competitively eluted by addition of 25 µg target protein. Eluted phages were amplified by incubation with ER2738 Escherichia coli cells (New England Biolabs) at 37 °C for 4.5 hours. Amplified phages were precipitated using PEG/NaCl and titrated. The second and third selection rounds were carried out in an identical way, using approximately 2×10^{11} of the phages eluted and amplified in the previous round.

Illumina MiSeq HTS

Amplified phage eluates from each of the 3 selection rounds were subjected to PCR using the Phusion High-Fidelity DNA polymerase (New England Biolabs). To mitigate PCR bias, the PCR reactions were carried out in triplicates and the number of cycles was limited to 20. The PCR was carried out as recommended by the manufacturer, using 25 µL reactions with 1 µL 1:10 diluted phage eluates as template and an annealing temperature of 60 °C. For sequencing of the phage library, the template was 1 µL of a 1:10 diluted library. The variable part of the phage library was specifically amplified using primers designed for Illumina sequencing (P5-: 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCGAA-TTCCTTTAGTGGT P7-: 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTAAACAACTTTCAACAGT). The

corresponding PCR fragment was extracted from a 2% agarose gel using the Qiaquick Gel Extraction Kit (Qiagen). The triplicate PCR reactions were pooled before gel extractions were carried out. The DNA concentration was measured on a Qubit fluorometer (Life) and the amplicon size was validated on a 2100 Bioanalyzer (Agilent). Next, sequencing adaptors coupled with barcodes (the Nextera Index Kit PCR primers 5'AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCA-GCGTC, 5'CAAGCAGAAGACGGCATAC-GAGAT[i7]GTCTCGTGGGCTCGG) were ligated onto the DNA fragments at the DTU Multi-Assay Core (DMAC) facility. The samples were pooled and sequenced on the Illumina MiSeq platform using paired-end 2x150 cycles, as described by the manufacturer.

Converting sequencing reads to peptides

DNA reads obtained from the sequencing were filtered automatically by sample barcodes. The paired-end reads were merged using the FLASh program (http://ccb.jhu.edu/software/FLASH/) using default values except for max_mismatch_density, where the stringency was increased to 0.1. The DNA reads were then translated in all six reading frames and aligned to the protein regions of the phage protein that flank the displayed 7-mer peptide using the BLAST algorithm. Then 7-mer peptides and their originating DNA sequences were extracted from reads that successfully resulted in alignments where the protein sequence aligned to both flanking regions with a 7-mer peptide gap. The peptides were further processed by i) removing peptides where at least 1 DNA base pair (bp) coding for the peptide were assigned a Phred quality score below 20 and ii) removing reads where the DNA sequence did not match the NNK-codon pattern intrinsic to the Ph.D.-7™ phage library (N = any nucleotide, K = G or T).

Aligning peptides to the target sequence

Peptides that were selected from the phage libraries were mapped to the target protein by aligning each peptide to overlapping 7-mers from the primary sequence using the same alignment parameters as in Chapter 1. Each phage peptide was assigned the starting position of the overlapping 7-mer that provided the best alignment score. Alignments with a score below zero were discarded. The significance of alignment scores was assessed by aligning 1 million random peptides to each overlapping 7-mer and reporting the 99.99 percentile of the alignment scores.

Results

Sequencing output

Sera from 3 animals that were immunized with the target protein were used in a set of phage display experiments. Serum was obtained before immunization, thus acting as controls and establishing a baseline, as well as 19 days after the first immunization where target-specific antibodies had developed. Two animals (1 and 6) were categorized as “high responders”, since their serum showed good binding capacity in ELISA at Day 19. The remaining animal (animal 2) was categorized as a “low responder” due to a lower signal on ELISA. Both groups were included to test whether there were any notable differences in epitope patterns between high and low responders. The phage particles were selected over 3 rounds for their interaction with serum and eluted with the target protein. For all 6 samples and all selection rounds the eluted phages were amplified and sequenced by MiSeq HTS. In addition, a subset of the Ph.D.-7™ phage library was also sequenced. Consequently, 19 samples were pooled and sequenced in total.

The sequencing output and quality was high. In total, about 20,600,000 raw reads were obtained with a relatively even distribution among the 19 barcodes. 91% of the identified bases were assigned a Phred quality score > 30, indicating a high quality of the majority of the sequenced base pairs. Consequently, a high number of peptides were assigned to each sample after sorting and alignment. Specifically 18,106,337 peptide sequences passed the quality filters and were assigned to a sample. This represented an average of 952,965 peptides per sample (range 416,578-1,461,630).

The phage selection process

As was the case in previous studies (Chapter 1 & 2), it was possible to use the HTS data to monitor the phage selection process from selection round 1 through 3, see Figure 2. The number of unique peptides decreased with each selection round. In the first selection round an average of 7,410 unique peptides were obtained, dropping to 2,655 and finally 985 unique peptides after the second and third selection round, respectively. When the combined frequencies were examined, they revealed a convergence of peptides, where the most frequent peptides were taking up a greater portion of the total number of peptides (Figure 2B). An example of the development can be observed in Figure 3, which shows the ten

most frequent peptides in a sample derived from an animal after immunization. Tables with the ten most frequent peptides for all samples can be found in Appendix B.

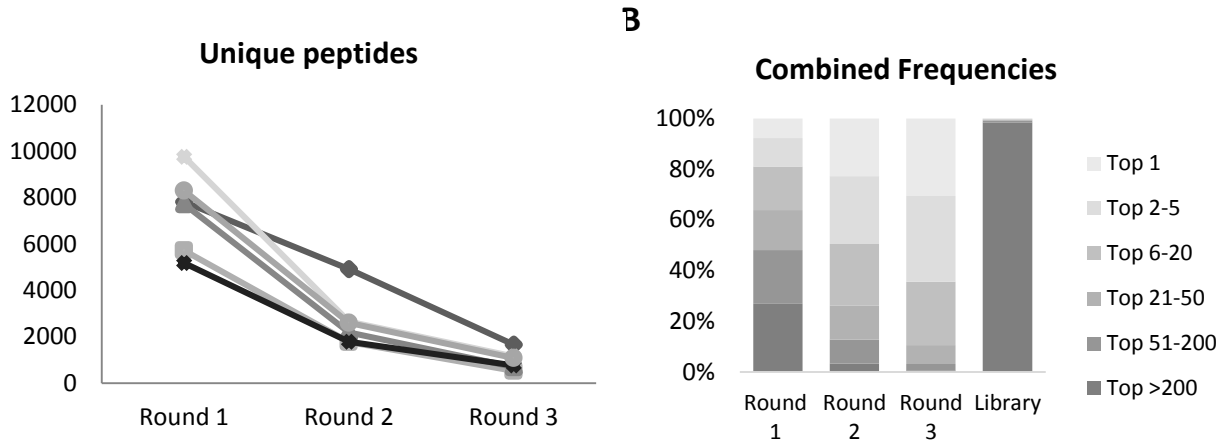


Figure 2. The phage selection process. **A)** A plot of the number of unique peptides identified in each sample for selection round 1-3. **B)** A stacked bar chart showing the combined frequency of certain rank intervals. Specifically, the frequency of the most frequent peptide (designated "Top1") along with the combined frequencies of the peptides ranked 2–5, 6–20, 21–50, 51–200 and below 200. The average of the combined frequencies for all samples is shown

A

Round 1		Round 2		Round 3	
TMTVYTK	17.0%	TMTVYTK	17.5%	TVNFKLY	37.6%
SATWNPV	3.6%	TVNFKLY	14.9%	QLAVAPS	10.2%
SHENFTS	3.2%	QLAVAPS	8.2%	DSSLFAL	8.4%
YTNHQPQ	2.7%	DSSLFAL	6.4%	YGAKNNL	7.4%
QEVVTNH	2.0%	SATWNPV	5.2%	TMTVYTK	6.5%
QLAVAPS	1.4%	YTNHQPQ	4.5%	GPLHAQF	4.7%
TSAHHRM	1.3%	YGAKNNL	3.2%	SAHMRTM	2.3%
VSGFRID	1.2%	SHENFTS	3.0%	HYIDFRW	2.2%
TVNFKLY	1.1%	TSAHHRM	2.7%	ESRVMSR	2.0%
NTQLHPS	1.0%	GTSIYLH	2.4%	TSAHHRM	1.6%

B

Library		
Sequence	**Reads**	**Frequency**
QLYREFN	380	0.07%
TTQVLEA	272	0.05%
DAIPTSV	249	0.05%
TVNFKLY	231	0.04%
KMISATE	196	0.04%
YPWFIRA	154	0.03%
WGRISHV	145	0.03%
HSHTLTW	126	0.02%
IDRTQFM	119	0.02%
WSLSELH	112	0.02%

Figure 3. A) A table with the ten most frequent peptides in selection round 1-3 against a sample (specifically 2B) after immunization is shown. **B)** A table with the ten most frequent peptides for the sequenced phage library, including the number of sequencing reads for each peptide. A total of 537,495 reads were split among 390,098 unique peptides.

Sequencing the phage library

A portion of the phage library was sequenced along with the samples to examine the library diversity. As expected, the output from the library was a lot more diverse than what was obtained from the samples (see Figure 2B and 3B). A total of 537,495 peptide reads were split among 390,098 unique peptides. Of the unique peptides, 292,516 were only represented by a single sequencing read. Even though the diversity was high, there were some peptides that were identified many times (Figure 3B). While the combined frequency of the ten most frequent peptides was just 0.37 %, they were still highly prevalent compared to the majority of the unique peptides.

An examination of the frequent library peptides revealed an overlap with peptides that had previously been identified with high frequency in other phage selections. In Chapter 1, 82 TUP candidates that emerged in >2 independent samples, including control samples, were identified. Of the 82 candidates, 33 of them were among the 100 most frequent library peptides. Especially the ten most frequent library peptides were ubiquitously detected: Nine were among the 82 candidates and 6 were among the ten most dominant TUP candidates, identified in > 5 independent samples. The most frequent peptide in the library, QLYREFN, was also the most widely recognized TUP candidate, as it had been identified in 11 out of 12 independent samples. There were also a considerable overlap between the top 10 lists in the present study and the frequent library peptides (Appendix B).

Aligning the phage peptides to the target sequence

In order to investigate whether the selected peptides matched the primary target protein sequence, thus suggesting a linear epitope, all peptides were aligned to the primary sequence of the target (Figure 4). This was inspired by previous successful results using this approach (Chapter 1). While there were singular peptides in the immunized samples that were above background, the pattern did not appear considerably different from the baseline samples. There was no convergence of positive peptides to a specific position, thus suggesting a selection of related peptides with similarity to the primary target sequence. Thus, no linear epitopes could convincingly be identified by this approach.

An attempt was made at clustering peptides from all immunized samples and contrasting to the peptides in the baseline samples, similar to the clustering approach applied in Figure 4 in Chapter 1. However, the main clusters that emerged included peptides from the baseline samples, indicating that

they were TUP-derived motifs (data now shown). Further optimization of the analysis may be needed to determine whether there are any relevant clusters in the samples.

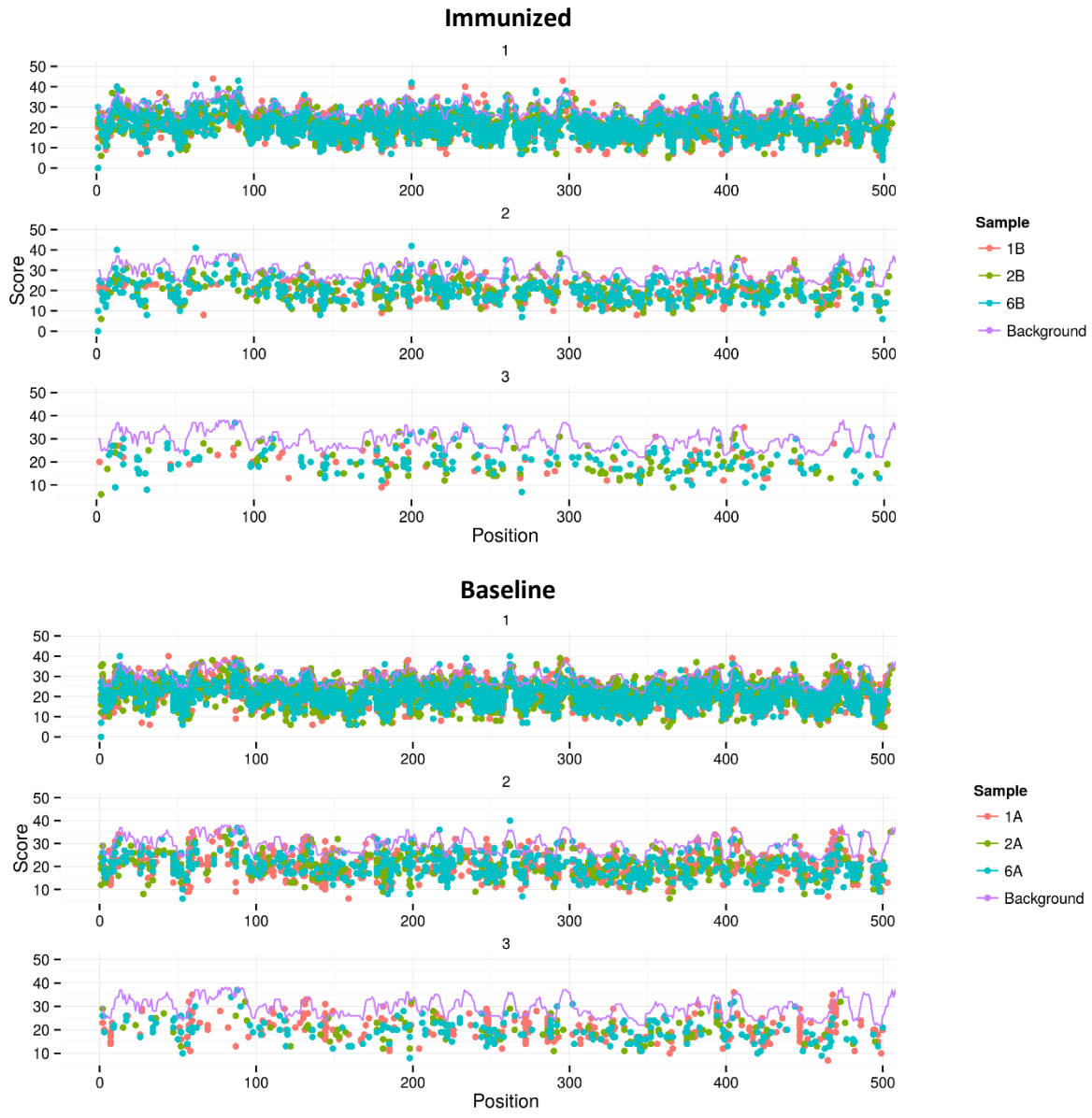


Figure 4. Alignment of immunized-derived peptides (top) or baseline-derived peptides (bottom) to the primary sequence of the target. Peptides were mapped by aligning each phage-derived peptide to overlapping 7-mers from the primary sequence using pairwise alignment. Each phage peptide was plotted at the starting position of the overlapping 7-mer that provided the best alignment score. The top row is a plot of the peptides from selection round 1, while the middle and bottom row are for selection round 2 and 3, respectively. The purple line represents a significance assessment which is the 99.99 percentile alignment scores obtained after aligning 1 million random peptides to each overlapping 7mer. Alignments with a score below zero were discarded. The figure has been cropped around position 500.

Discussion

In this project, animals were immunized with a potential therapeutic target and sera were obtained before and after immunization. Phage libraries were selected against the sera in an attempt to map the target epitopes. Switching from an IonTorrent to a MiSeq platform greatly increased the sequence output and quality. E.g. the number of reads increased from about 1.5 million to about 20 million. The peptide output from each selection round enabled an examination of the selection process. Finally, the peptides were aligned to the target sequence in an attempt to identify linear epitope candidates; however, no such candidates could be positively identified.

The overall phage selection process appeared to be similar to what was previously observed in Chapter 1. However, a higher number of unique peptides were observed. This can be explained by at least two effects. The MiSeq platform notably increased the sequence output and it is likely that an increased sequencing depth lead to a more complete sequencing of the eluted phage library. The other factor that likely also increased the number of peptides was the fact that the sequences in this study had not been filtered for sequences derived from highly prevalent sequences, which was the case in the previous study. Due to the increase in data quality, it can be expected that the prevalent sequences will give rise to fewer derivatives; however, the increased sequencing output should result in additional derivative peptides. When the combined frequencies were examined (Figure 2B) they exhibited similar trends to previous experiments. There was a tendency towards less convergence to a single peptide. This may be explained by differences in the phage experiments, the PCR amplification (where better performing primers led to a reduction from 25 to 20 cycles) as well as the change in sequencing platform.

Sequencing of the phage library provided insights into its diversity and the nature of TUPs. It was confirmed that the library had a very high diversity. The great overlap between frequent library peptides and peptides that were highly frequent in this study, along with previously suggested TUP candidates, indicated that many TUPs are prevalent in the initial library. This indicates that these TUPs have a proliferation advantage rather than a selection advantage since the original library has not undergone selection. It is important to note that the prevalence of these proliferation-TUPs in the selected phage eluates may be overestimated since the phages are sequenced after amplification in bacteria. The overlap between frequent library peptides and TUPs identified in independent phage screenings suggest that knowledge of the phage library may assist in identification of putative TUPs. Notably, Matochko *et al.* also sequenced the Ph.D.-7™ phage library before and after amplification, however no overlap between the top 100 peptides identified in this study and the 30 most frequent peptides before/after

amplification in the Matochko study was observed. This could be explained by differences in sample preparation and sequencing setups but also suggests that the frequent peptides may change between different batches of the same phage library. Altogether, sequencing the phage library in parallel to phage selection rounds seems to be a valid approach to identify putative TUP sequences.

One could expect that the increased antibody levels in immunized animals would make epitope mapping more straightforward than mapping based on IgE in human sera (Chapter 1). Nonetheless, an initial attempt at aligning all peptides did not identify distinct linear epitope candidates. Correspondingly, no relevant motifs emerged when an attempt at clustering the peptides was performed. These were the approaches we previously employed to successfully identify a linear epitope[5]. Further bioinformatic analysis is required to elucidate whether any putative epitopes can be identified in the data. The most obvious approach would be to further examine the data for peptide clusters that are unique to the immunized animals, possibly without clustering across different samples. Clusters may help identify epitope candidates. If such peptide clusters did not align with high alignment scores to the primary target protein, it may indicate that the peptides mimic a conformational epitope. Furthermore, investigating whether prevalent peptides have significantly increased alignment scores is another approach. It is also possible that the selection procedure did not work as intended and that further optimization of the phage experiments are required. In summary, while the attempt to map the epitopes in sera from immunized animals worked at the technical level, further studies will determine whether any epitopes can be identified in the data.

Contributions

Nikolaj Dietrich at Symphogen A/S contributed to the design of the study. Animal immunization and sample isolation was carried out by Symphogen A/S. Christian Skjødt Hansen made adjustments to the bioinformatic pipeline in order to incorporate the MiSeq data and carried out the alignment to the target protein. Mireia Solá participated in the phage display experiments. Staff at the DTU Multi-Assay Core (DMAC) facility performed the MiSeq sequencing. Anders Christiansen designed the study and performed the phage display experiments, the sample preparation as well as the data analysis.

References

1. Beck, A., Wurch, T., Bailly, C. & Corvaia, N. Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.* **10,** 345–52 (2010).

2. Brekke, O. H. & Løset, G. Å. New technologies in therapeutic antibody development. *Curr. Opin. Pharmacol.* **3,** 544–50 (2003).

3. Liu, J. K. H. The history of monoclonal antibody development – Progress, remaining challenges and future innovations. *Ann. Med. Surg.* **3,** 113–116 (2014).

4. Sharon, J., Rynkiewicz, M. J., Lu, Z. & Yang, C.-Y. Discovery of protective B-cell epitopes for development of antimicrobial vaccines and antibody therapeutics. *Immunology* **142,** 1–23 (2014).

5. Christiansen, A. *et al.* High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5,** 12913 (2015).

Chapter 4

High-resolution epitope mapping of peanut allergens demonstrates individualized and highly persistent antibody binding patterns at the residue level

High-resolution epitope mapping of peanut allergens demonstrates individualized and highly persistent antibody binding patterns at the residue level

Christian S. Hansen[1], Martin Dufva[2]*, Katrine L. Bøgh[3], Eric Sullivan[4], Jigar Patel[4], Thomas Eiwegger[5,6], Zsolt Szépfalusi[5], Morten Nielsen[1,7], Anders Christiansen[2]

[1] Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Department of Micro- and Nanotechnology, Technical University of Denmark, Kgs. Lyngby, Denmark,

[3] National Food Institute, Technical University of Denmark, Søborg, Denmark

[4] Roche NimbleGen, Madison, Wisconsin, the United States of America

[5] Department of Pediatrics, Medical University of Vienna, Vienna, Austria

[6] Department of Pediatrics, Hospital for Sick Children, University of Toronto, Toronto, Canada

[7] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

* Correspondence and requests for materials should be addressed to Martin Dufva (Dufva@nanotech.dtu.dk)

Abstract

Background:

Characterization of allergen epitopes is an essential task in unraveling the molecular mechanisms of allergy. One way to investigate epitope patterns is by peptide microarrays. Recent advances in peptide array technology have greatly increased the throughput thus enabling the study of a vast number of peptides in parallel.

Objective:

We sought to use high-density peptide microarrays to characterize B cell epitopes at high resolution, including information on the importance of specific residues for antibody binding.

Method:

Overlapping 12-mer peptides, with a single offset, covering the peanut allergens Ara h 1, 2, 3, 6, 8 and 9 were synthesized on high-density peptide microarrays. Furthermore, to determine the importance of specific residues for antibody binding, a complete alanine scan was included. About 20,000 peptides were synthesized to accomplish this. The IgE and IgG4 reactivity was measured for 7 patients with severe peanut allergy that were sampled at multiple time points over a course of 4-10 years.

Results:

Patients had heterogeneous epitope patterns that converged on certain epitope hotspot areas. At the hotspot locations, alanine substitutions revealed both individualized residue-level binding patterns that differed among patients as well as residue patterns that were shared across all investigated patients. The epitope patterns were highly persistent over time. IgE and IgG4 epitopes were highly co-localized.

Conclusion:

Next generation peptide microarrays enabled a high-resolution epitope analysis and revealed that patients react differently to well-described epitope hotspots. The epitope persistence implies that peanut allergy is a relatively static condition.

Key words: Peanut allergy, epitope, peptide microarray, epitope mapping, IgE, IgG4, allergy persistence.

Introduction

Peanut allergy is considered a major food allergy due to its prevalence, persistence and its association with severe symptoms[1,2]. Previous studies suggest that clinical reactivity to allergens, including peanut, might be related to allergen-specific IgE epitope patterns, diversity, and avidity, all of which likely play a role in the effect of IgE on basophils and mast cells[3–6].

Numerous peanut allergens (designated Ara h 1-Ara h 15)[7] have been characterized, however, several studies have shown in both patients with peanut allergy and sensitized animals that the majority of the IgE response is directed to Ara h 1, Ara h 2, Ara h 3, and Ara h 6[3,8,9]. In addition, a relatively high frequency of Ara h 9 reactivity has been observed in Southern Europe[10]. Finally, Ara h 8 is likely responsible for cross-reactivity to birch pollen allergy[11]. For several of the major allergens, IgE-binding epitopes have been characterized[12–15].

The epitope binding patterns for some of the peanut allergens have previously been studied using peptide microarrays[4,16–20]. On one hand, these studies have demonstrated a pronounced heterogeneity in the epitope binding patterns between individual allergic patients. On the other hand, they have shown that antibody binding tend to converge on some areas of the allergen sequence, sometimes designated as epitope hotspots. This observation of recurrent epitopes has inspired the search for epitope biomarkers[13,18,20]. Several studies have reported a correlation between the number of recognized epitopes (epitope diversity) and the severity of the disease in allergic patients[4,17,18]. One study observed the development of new epitope recognition patterns following the onset of oral immunotherapy[19].

The function of allergen-specific IgE appears to be modified by the presence of IgG4[21]. Elevated IgG4 levels have been associated with a protective effect. This effect includes a suppression of hypersensitivity reactions[22–24] as well as a decreased risk of developing peanut allergy in infants subjected to dietary intake of peanuts[25,26]. Along these lines, the desensitization that follows peanut immunotherapy has been associated with increased IgG4 levels[19,27,28].

New developments in peptide array technology have enabled the studies of up to 2.1 million peptides in parallel[29,30]. We utilize these advances to perform a detailed epitope mapping of all major peanut allergens (Ara h 1, 2, 3, 6, 8 and 9), including a full alanine scan of all proteins. This provides an opportunity to study epitopes at an exceptionally high resolution and determine the importance of each

residue for antibody binding. We observe that while patients on the overall level appear to react similarly to a given epitope, their binding patterns are heterogeneous when the epitopes are examined at the amino acid level. Furthermore, there is a high correlation between IgE and IgG4 binding patterns possibly indicating a common clonal origin. Finally, we demonstrate that these epitope binding patterns are highly stable over multiple years.

Methods

Study population

Patients with peanut allergy were recruited at the Department of Pediatrics, Medical University of Vienna, Austria. In total, 17 samples obtained at different time points from 7 peanut allergic adults (P1-P7) were analyzed. This corresponded to 2 or 3 samples from each patient, which covered an average time interval of 8.3 years (range 4-10 years). The samples were generally from adult patients (aged 18-40). One exception was P2, a child aged 8-12. In addition, sera from 5 peanut-tolerant subjects (C1-5) were included as controls. No peanut-specific IgE reactivity was observed for the control samples. The peanut allergic patients were diagnosed by an experienced allergologist and had a convincing history of peanut-related symptoms. The clinical history for each patient can be found in Table E1 in Appendix C. The protocol was approved by the local Ethical Committee of the Medical University of Vienna and written informed consent was obtained (protocol no. EK 428/2008).

High-density peptide microarrays

A peptide library consisting of single-offset overlapping 12-mer peptides covering the primary sequence of Ara h 1, 2, 3, 6, 8 and 9 in triplicates was generated. Furthermore, the library consisted of 1 non-redundant copy of all single-residue alanine derivatives of the native peptides yielding a total allergen peptide library of 22,490 peptides. Serving as control peptides, the library included 50,582 peptides deriving from peanut-unrelated proteins, specifically 148 snake venom toxins. Peptide synthesis was accomplished as described previously[31] by *in-situ* photolithographic synthesis in a Roche-NimbleGen Maskless Array Synthesizer using an amino functionalized substrate carrying a photosensitive protection group.

IgE and IgG4 Immunoassay

The final microarrays were incubated overnight at 4 °C with individual samples mixed with binding buffer in a 1:2 dilution (for serum samples) or 1:4 dilution (for plasma samples) in a final volume of 20 µL. This was followed by incubation with goat anti-Human IgE conjugated with tetramethylrhodamine (Life, A18798) and mouse anti-Human IgG4 conjugated with Alexa Fluor® 647 (SouthernBiotech, 9200-31) at room temperature for 3 hours. Finally, the arrays were scanned using a MS200 microarray scanner at 532 nm (IgE) and 635 nm (IgG4) excitation and signals were extracted using NimbleGen DEVA signal extraction software.

Data analysis

The fluorescence intensity of all recorded peptide fields was mapped back on the originating peanut allergen and indexed as the position of the first peptide residue in the allergen. To filter away potential microarray artifacts, the median intensity was used for all triplicate signals. Next, a rolling median of the peptides with a window size of 3 was applied; effectively eliminating singular peptide responses. A normalization step was implemented in order to compare signals across multiple microarray wells. First, all peptide signals were normalized according to the baseline signal from the non-relevant peptides derived from snake toxins (see above). This was accomplished for each well and antibody subtype by iteratively calculating the baseline mean of the non-relevant peptides, removing outliers more than 2 standard deviations away until a constant mean, m, was obtained. Next, a standard deviation, s, was calculated on all non-relevant peptides with a signal lower than the baseline mean. Finally, a signal-to-baseline Z-score was calculated for each peanut peptide signal, x, using $Z = (x - m)/s$. To correct for peptides consistently showing a signal above baseline, the Z-score was subtracted by the maximum Z-score of identical peptides probed with control serum. The control sera were from subjects with no detected IgE reactivity and no epitopes (peaks) were observed in the raw data. The process of normalization and correction is illustrated on actual data in Figure E2 in Appendix C. A cutoff of two standard deviations ($Z > 2$) was used to define positive IgE and IgG4 binding peptides. This cutoff was used for i) calculation of percentage of allergen sequence with positive binding peptides, ii) percentage of patients sharing positive binding peptides and iii) for determining which peptides should be included in the alanine-substitution analysis.

Alanine-substitution analysis

A complete alanine scan of all native peanut-originating peptides was included on the microarray. Each substituted peptide was expressed as the percentage loss of signal intensity relative to the corresponding native peptide. To report the substitution effect, the median percentage loss in signal of overlapping peptides substituted within the same allergen residue was calculated. Only substitutions in positive native peptides (see above) were included in the median calculation. For comparison of the substitution effect of epitope binding residues between samples, a Spearman rank-order correlation coefficient was calculated on the median substitution effect of residues from one sample, with those of another, given the value was present in both samples.

Results

High-density peptide microarrays were used to make a complete map of linear IgE and IgG4 epitopes for the six peanut allergens Ara h 1, 2, 3, 6, 8 and 9. Overlapping 12-mer peptides with a single offset covering the sequence of the peanut allergens were *in situ* synthesized in triplicates (4,713 peptides). Furthermore, for each individual peptide all amino acids were substituted, one at a time, with an alanine to be used for substitution analysis (17,777 peptides). In total, 17 samples were analyzed from 7 peanut allergic patients. 2 or 3 samples covering an average time interval of 8.3 years (range 4-10 years) were obtained from each patient. In addition, sera from 5 subjects with no reactivity to peanuts were included as controls.

Patient serum IgE and IgG4 antibodies selectively interacts with allergen peptides

In general, sera from peanut allergic patients had a much higher IgE and IgG4 reactivity against peanut allergens than control subjects (Figure 1A and 1B). As can be observed in Figure 1A+B, IgG4 signals displayed a higher degree of baseline variations than IgE. All signal values were therefore normalized by a Z-score based on the baseline of non-relevant peptides in each of the microarray wells and corrected for unspecific binding, in which identical peptides consistently had a signal above baseline (Figure 1C). After these corrections, a total of 1463 (6.3%) IgE-positive and 1379 (6.0%) IgG4-positive peptides ($Z > 2$) were identified among the patients. In comparison, 1 (0.01%) IgE-positive and 177 (2.14%) positive IgG4 peptides were identified among the control subjects.

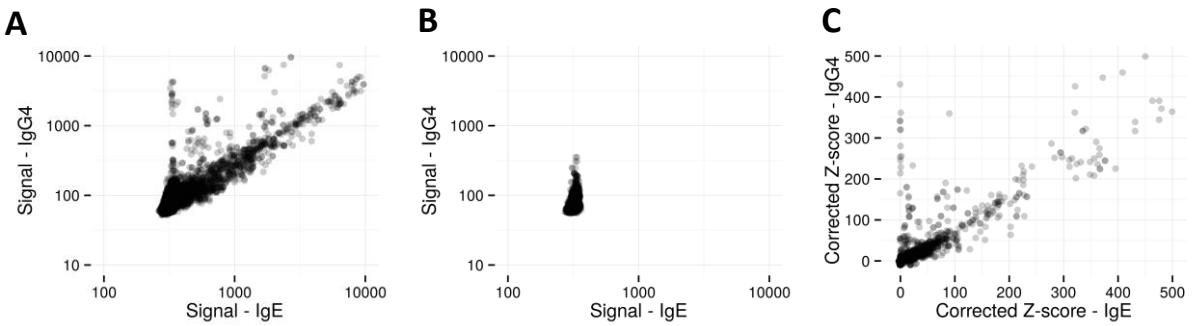


Figure 1. Scatterplots of IgE versus IgG4 signal. **(A)** Signal from peanut-peptides probed with peanut allergic patient samples (N = 25,452). **(B)** Signal from peanut-peptides probed with control samples with no peanut reactivity (N = 9,090). **(C)** Like (A) but depicting the corrected Z-score value for each peanut-peptide. Outlier filtration has been applied on triplicate peptides, see Methods for details.

Ara h 1 and 2 had the highest number of epitopes of the investigated allergens

Overall 0-14.4% of the peptides from each allergen were positive ($Z > 2$) for antibody binding. Ara h 1 and 2 had the highest number of positive peptides with an average of 10.0% for IgE and 9.4% for IgG4 for Ara h 1 across all patients. These numbers were 14.4% for IgE and 14.0% for IgG4 for Ara h 2. The observation that Ara h 1 and 2 have the highest epitope density is consistent with previous studies[9,17]. One patient, P7, did not display any positive peptides for any allergen at both analyzed time points and was omitted from further analyses. One sample from P3 (P3_13) had unspecific IgG4 signal across all peptides and was excluded from the analyses.

Figure 2 illustrates the corrected Z-score of each native peptide for Ara h 1 and Ara h 2. The remaining allergens exhibited between 0% and 4.3% positive peptides and corresponding figures can be found in Figures E4 and E5 in Appendix C. Figure 2 revealed the position and approximate boundaries of the epitope regions. The epitope patterns were generally heterogenous among different patients. However, regions that were recognized by several patients were clearly present, such as region 85-95 in Ara h 1 in Figure 2. For Ara h 1 34.6% (IgE) and 28.8% (IgG4) of the binding peptides were shared between three or more patients while for Ara h 2 the percentage were 48.4% (IgE) and 48.3% (IgG4). For Ara h 3 and 6 the fraction of binding peptides shared between three or more patients ranged between 6.8% and 9.7% while binding peptides from Ara h 8 and 9 were infrequent and they were not shared among patients. See Figure E3 in Appendix C for a complete overview of shared peptides.

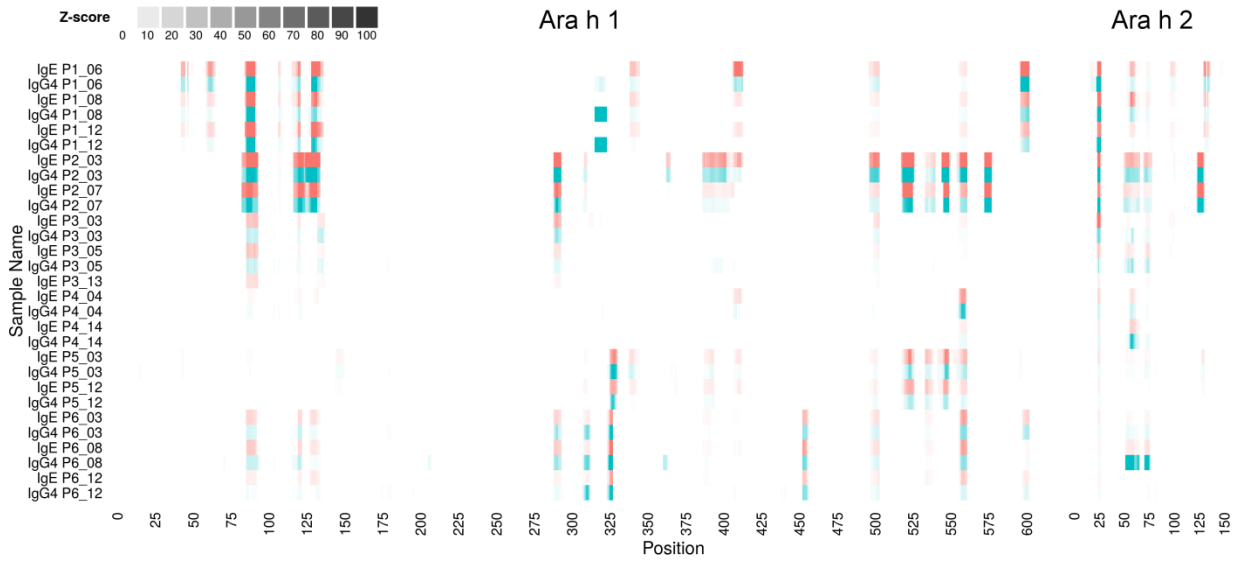


Figure 2. Corrected Z-score of native peptides originating from Ara h 1 and Ara h 2. The Z-score is reported at the starting position of a given peptide in the allergen sequence (x-axis) for both IgE (red) and IgG4 (blue). Each row indicates IgE or IgG4 channels from patient samples P1 through P6 at a certain year of sampling (e.g. P1_06 is Patient 1 sampled in 2006). Similar figures for the remaining allergens can be found in E4 and E5 in Appendix C.

IgE and IgG4 responses were highly correlated across all patients

Peptides that were positive for an IgE signal were often found to also be positive for an IgG4 signal. On average, 68.6% of the positive binding peptides were both IgE and IgG4 positive. Of these peptides, the corrected Z-scores of IgE and IgG4 signals were highly correlated (Spearman's rho = 0.90 across all 6 allergens), as indicated in Figure 1C. The specificity of the anti-IgE and anti-IgG4 secondary antibodies was confirmed by ELISA (data not shown).

The epitope patterns were highly persistent over multiple years

The results of the alanine substitution scans of all native peptides of Ara h 1 are summarized in Figure 3A. Similar figures can be found for the remaining allergens in Figure E6-E8 in Appendix C. The overall epitope pattern corresponds to the pattern shown for the native peptides in Figure 2. Within the different patients, the epitope patterns were highly persistent over multiple years, whereas larger differences were observed between patients. This observation corresponds to the higher Spearman's rank correlation of the substitution effects (Figure 3B and 3C) for samples obtained from the same

patient (average 0.85 for IgE and 0.78 for IgG4) compared to the correlation between different patients (average 0.52 for IgE and 0.48 for IgG4). A detailed heatmap of correlations between samples can be found in Figure E9 in Appendix C.

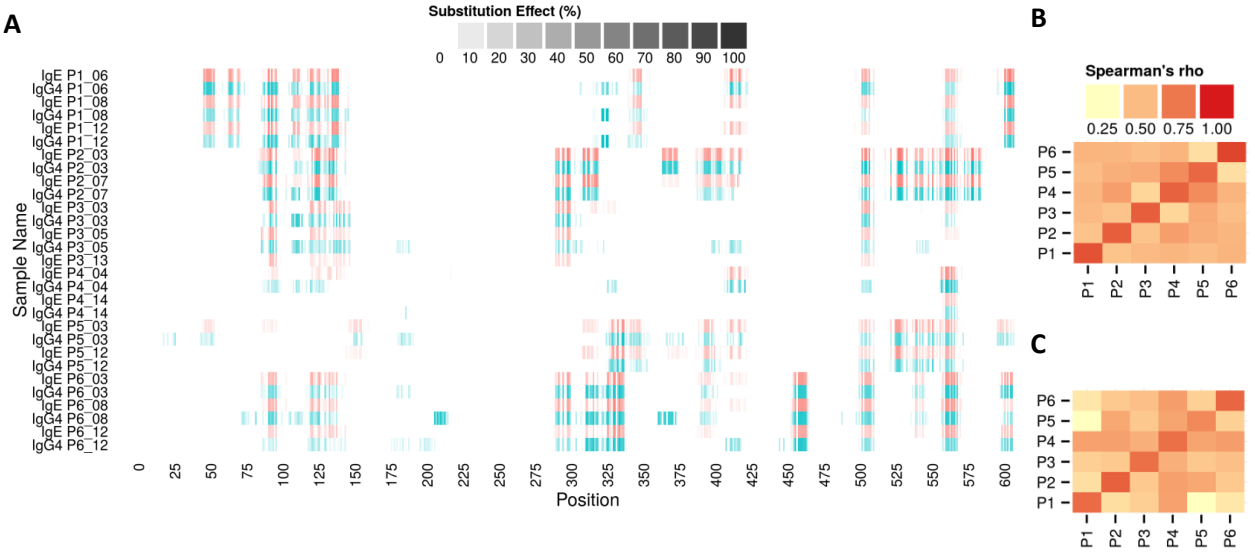


Figure 3. **(A)** Alanine substitution effect in Ara h 1 for residue-level epitope mapping. The relative effect of substituting a given residue (x-axis) to an alanine is reported as 0% to 100% loss of signal compared to the signal of the native 12-mer peptide (see the Methods section for details). Each row indicates IgE or IgG4 channels from patient samples P1 through P6 at a certain year of sampling (e.g. P1_06 is Patient 1 sampled in 2006). For figures of the remaining allergens, see Figures E6-E8 in Appendix C. **(B, C)** Correlation heatmaps of Spearman's rank correlation of substitution effects of allergen residues for IgE (B) and IgG4 (C) between patients. Only substitutions of positive peptides in both samples are included. For detailed sample-specific correlation, see Figure E9 in Appendix C.

Patients have individualized binding patterns at epitope hotspots

Although many binding peptides are shared among patients, as shown in Figure 2 and 3, the individual residues involved in the epitope are not necessarily the same (Figure 3). This is illustrated in greater detail in Figure 4, which shows an Ara h 1 epitope that was shared among all patients. An examination of the substitution effects showed that epitope regions, which are shared among patients, often have distinct binding profiles when examined at the residue level. Whereas some residues are critical for binding across all patients (an alanine mutation leads to a complete loss of binding), e.g. 99D in Figure 4, there are other residues that are of importance in some patients but irrelevant in others. Other epitope hotspots display a higher degree of similarity of the epitope binding patterns between patients, e.g. the

region 30-39 of Ara h 2 in which the residues DRR-Q-Q-ER are consistently involved in the epitope in the majority of patients (Figure E10 in Appendix C).

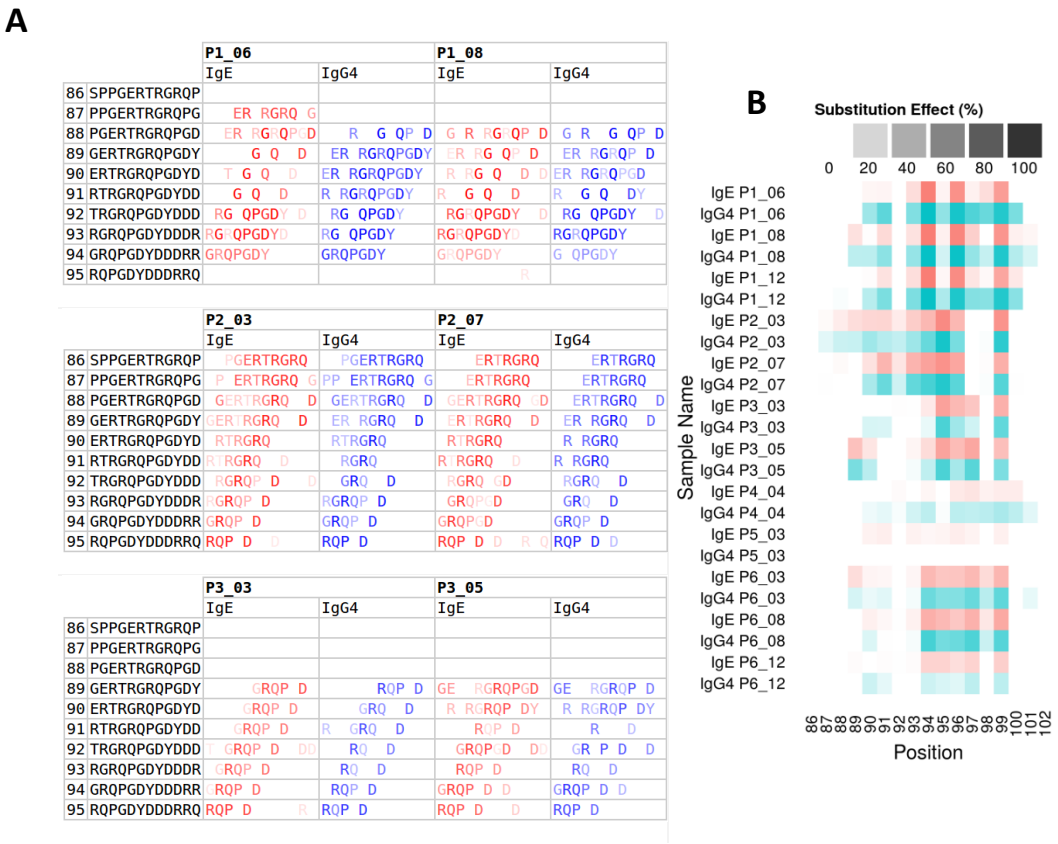


Figure 4. Alanine substitution effect for a sample epitope in Ara h 1 shown at high-resolution. The effect for IgE is shown in read and the effect for IgG4 in blue. **(A)** Alanine substitution analysis of 10 individual peptides spanning the 86-107 region of Ara h 1. Rows indicate each sliding window of overlapping 12-mer peptides undergoing substitution. Columns indicate patients P1, P2 and P3 as well as the year of sampling, e.g. P1_06 indicates Patient P6 sampled in 2006. Peptide residues are shown in a zero to full intensity scale representing a 0-100% decrease in signal upon alanine substitution, see methods section for details. **(B)** Median substitution effects of Ara h 1 protein residues 86-102 represented in multiple overlapping 12-mer peptides. See the Figure 3 legend for details.

Discussion

In this study, we applied a new generation of ultrahigh-density peptide microarrays to perform high-resolution linear epitope mapping of the major peanut allergens. The high resolution as well as the alanine-scan enabled residue-level insight into the linear IgE and IgG4 epitope binding patterns of the major peanut allergens. The level of detail made it possible to determine that i) while patients may react

to the same epitope location or hotspot, the individual residues involved in antibody binding are often divergent between patients, ii) the IgE and IgG4 binding patterns were highly correlated and iii) the epitope patterns of individuals were persistent over multiple years.

Examining the overall allergen reactivity patterns, epitopes were identified most commonly in Ara h 1, 2 and 3, amounting to 95.3% of all positive peptides. These observations were similar to findings in previous studies[4,9,17]. We observed that several patients reacted to the same areas in each of the major peanut allergens Ara h 1, 2 and 3. These areas corresponded well with the location of previously described epitopes and likely represent epitope hotspots[4,17,18]. However, high-resolution alanine scan revealed another layer of complexity concerning the binding patterns at the residue level. Patients had individualized binding patterns: whereas, for a given epitope-region some residues are consistently important for binding across all patients, there are other residues that are of great importance in some patients but irrelevant in others. This finding is supported by an early study by Stanley *et al.*, who made similar observations for selected Ara h 2 epitopes[14]. These findings demonstrate an individualized allergy response where epitope hotspots are recognized differently by individual patients. However, the fact that we identify residues binding patterns that are shared across all patients indicate that it may be feasible to identify critical residues that are always involved in antibody binding within a given hotspot region. Such knowledge may e.g. assist attempts at designing recombinant allergens for therapy[32] or correlating specific binding patterns to clinical outcome such as severity, persistence or treatment response. In order to elucidate which residues in the allergens that are critical for antibody binding in all patients, further studies with an increased number of subjects would be required.

The IgE and IgG4 epitopes very often co-localized. IgE and IgG4 epitopes have previously been reported to correlate[16,17,19]. Specifically, it has been reported that 30-50% of the peptides that were positive for IgG4 binding were also positive for IgE. However, studies have also described how the IgG4 signal was less pronounced than the IgE signal in peanut allergic patients[16–18]. The data presented here indicate that co-localization is more prominent that were previously suggested and that the IgG4 signal is comparable to the IgE signal. However, it should be noted that intensity is difficult to compare between studies since it is dependent on a number of factors including affinity and concentration for both primary and secondary antibodies as well as the applied detection system.

Previous studies of IgG4 in peanut oral immunotherapy have reported that there is a switch from IgE to IgG4 antibodies[19,27,28,33]. The hypothesis is that the emerging IgG4 antibodies have a protective "blocking" effect[21,26,34]. Our findings indicate that IgG4 is regularly present in the allergic response;

possibly suggesting that the sensitization obtained after immunotherapy is mainly gained by an increase in IgG4 levels or an increase in avidity instead of an induction of novel IgG4 specific antibodies. Recent studies of immunotherapy supports this hypothesis, since it has been shown that IgG4 is only induced to the specific allergens where pre-treatment specific IgE is present[35]. At the residue-level, the binding patterns were also very similar for IgE and IgG4 antibodies. This may indicate the antibodies are clonally related.

Since the patients included in the study were sampled at multiple time points, it was possible to investigate the epitope patterns over time. Overall, the epitopes were remarkably stable over a long period of up to 10 years. This is consistent with a previous observation by Flinterman *et al.* who reported that IgE and IgG4 epitope patterns were largely stable over a period of 20 months in children[17]. Alanine substitution effects of allergen residues were highly correlated within patients across multiple years, which indicate long-term clonal persistence of allergen-specific antibody-producing cells. Long-lived plasma cells that remain in the body long after exposure to allergen have previously been proposed to be responsible for this effect[36].

In conclusion, peanut allergy seems to be a highly stable condition characterized by epitope binding patterns that remain stable over many years. IgG4 binding correlated with IgE binding and was present in patients with allergy that had not been subjected to immunotherapy. The use of a new generation of peptide microarrays enabled the study of epitope patterns at a very high resolution. This established that patients react differently to the same epitope hotspot region; however, there are residues that seem to be important across patients. The new generation of peptide microarrays have a very high capacity enabling the study of >2,000,000 peptides per patient[29,30] which could be exploited for a number of applications. E.g. one could envision arrays that are used to get a detailed understanding of the relationship between IgE and IgG4 epitope binding patterns following allergy immunotherapy. Furthermore, arrays designed for high-resolution epitope mapping of all known allergens are within reach and could be useful to study e.g. cross-reactivity. Overall, the next generation of arrays appears to perform very well and may be an important tool for future studies within the allergy field and beyond.

Author contribution statements

C.S.H., M.D., K.L.B. and A.C. designed the study. C.S.H., E.S. and J.P. designed and performed peptide microarray experiments. C.S.H., M.D., M.N. and A.C. analyzed the data. T.E. and Z.S. provided and

characterized crucial reagents. C.S.H. and A.C. wrote the manuscript. All authors reviewed the manuscript.

References

1. Sampson, H. A. 9. Food allergy. *J. Allergy Clin. Immunol.* **111,** S540–7 (2003).

2. Burks, A. W. Peanut allergy. *Lancet* **371,** 1538–1546 (2008).

3. Lewis, S. A., Grimshaw, K. E. C., Warner, J. O. & Hourihane, J. O. B. The promiscuity of immunoglobulin E binding to peanut allergens, as determined by Western blotting, correlates with the severity of clinical symptoms. *Clin. Exp. Allergy* **35,** 767–773 (2005).

4. Shreffler, W. G., Beyer, K., Chu, T. H. T., Burks, A. W. & Sampson, H. A. Microarray immunoassay: Association of clinical history, in vitro IgE function, and heterogeneity of allergenic peanut epitopes. *J. Allergy Clin. Immunol.* **113,** 776–782 (2004).

5. Mitchell, A. J., Moss, N. D. & Collins, A. M. The biological activity of serum IgE changes over the course of a primary response. *Scand. J. Immunol.* **55,** 33–43 (2002).

6. Knol, E. F. Requirements for effective IgE cross-linking on mast cells and basophils. *Mol. Nutr. Food Res.* **50,** 620–4 (2006).

7. Allergen nomenclature approved by the World Health Organization and International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature Sub-committee. Available at: http://www.allergen.org. Accessed July 6, 2015.

8. Van Wijk, F., Hartgring, S., Koppelman, S. J., Pieters, R. & Knippels, L. M. J. Mixed antibody and T cell responses to peanut and the peanut allergens Ara h 1, Ara h 2, Ara h 3 and Ara h 6 in an oral sensitization model. *Clin. Exp. Allergy* **34,** 1422–8 (2004).

9. Koppelman, S. J., Wensing, M., Ertmann, M., Knulst, A. C. & Knol, E. F. Relevance of Ara h1, Ara h2 and Ara h3 in peanut-allergic patients, as determined by immunoglobulin E Western blotting, basophil-histamine release and intracutaneous testing: Ara h2 is the most important peanut allergen. *Clin. Exp. Allergy* **34,** 583–90 (2004).

10. Krause, S. *et al.* Lipid transfer protein (Ara h 9) as a new peanut allergen relevant for a Mediterranean allergic population. *J. Allergy Clin. Immunol.* **124,** 771–8.e5 (2009).

11. Mittag, D. *et al.* Ara h 8, a Bet v 1-homologous allergen from peanut, is a major allergen in patients with combined birch pollen and peanut allergy. *J. Allergy Clin. Immunol.* **114,** 1410–7 (2004).

12. Burks, A. W. *et al.* Mapping and mutational analysis of the IgE-binding epitopes on Ara h 1, a legume vicilin protein and a major allergen in peanut hypersensitivity. *Eur. J. Biochem.* **245,** 334–339 (1997).

13. Beyer, K. *et al.* Measurement of peptide-specific IgE as an additional tool in identifying patients with clinical reactivity to peanuts. *J. Allergy Clin. Immunol.* **112,** 202–7 (2003).

14. Stanley, J. S. *et al.* Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Ara h 2. *Arch. Biochem. Biophys.* **342,** 244–53 (1997).

15. Rabjohn, P. *et al.* Molecular cloning and epitope analysis of the peanut allergen Ara h 3. *J. Clin. Invest.* **103,** 535–542 (1999).

16. Shreffler, W. G., Lencer, D. a., Bardina, L. & Sampson, H. a. IgE and IgG4 epitope mapping by microarray immunoassay reveals the diversity of immune response to the peanut allergen, Ara h 2. *J. Allergy Clin. Immunol.* **116,** 893–899 (2005).

17. Flinterman, A. E. *et al.* Peanut epitopes for IgE and IgG4 in peanut-sensitized children in relation to severity of peanut allergy. *J. Allergy Clin. Immunol.* **121,** 737–743.e10 (2008).

18. Lin, J. *et al.* A bioinformatics approach to identify patients with symptomatic peanut allergy using peptide microarray immunoassay. *J. Allergy Clin. Immunol.* **129,** 1321–1328.e5 (2012).

19. Vickery, B. P. *et al.* Peanut oral immunotherapy modifies IgE and IgG4 responses to major peanut allergens. *J. Allergy Clin. Immunol.* **131,** 128–34.e1–3 (2013).

20. Otsu, K., Guo, R. & Dreskin, S. C. Epitope analysis of Ara h 2 and Ara h 6: characteristic patterns of IgE-binding fingerprints among individuals with similar clinical histories. *Clin. Exp. Allergy* **45,** 471–84 (2015).

21. Aalberse, R. C., Stapel, S. O., Schuurman, J. & Rispens, T. Immunoglobulin G4: an odd antibody. *Clin. Exp. Allergy* **39,** 469–77 (2009).

22. Hussain, R., Poindexter, R. W. & Ottesen, E. A. Control of allergic reactivity in human filariasis. Predominant localization of blocking antibody to the IgG4 subclass. *J. Immunol.* **148,** 2731–7 (1992).

23. Platts-Mills, T. A., Vaughan, J. W., Blumenthal, K., Pollart Squillace, S. & Sporik, R. B. Serum IgG and IgG4 antibodies to Fel d 1 among children exposed to 20 microg Fel d 1 at home: relevance of a nonallergic modified Th2 response. *Int. Arch. Allergy Immunol.* **124,** 126–9 (2001).

24. Michils, A. *et al.* Early effect of ultrarush venom immunotherapy on the IgG antibody response. *Allergy* **55,** 455–62 (2000).

25. Du Toit, G. *et al.* Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy. *N. Engl. J. Med.* **372,** 150223141105002 (2015).

26. Santos, A. F. *et al.* IgG4 inhibits peanut-induced basophil and mast cell activation in peanut-tolerant children sensitized to peanut major allergens. *J. Allergy Clin. Immunol.* **135,** 1249–56 (2015).

27. Kim, E. H. *et al.* Sublingual immunotherapy for peanut allergy: Clinical and immunologic evidence of desensitization. *J. Allergy Clin. Immunol.* **127,** 640–646.e1 (2011).

28. Jones, S. M. *et al.* Clinical efficacy and immune regulation with peanut oral immunotherapy. *J. Allergy Clin. Immunol.* **124,** 292–300, 300.e1–97 (2009).

29. Forsström, B. *et al.* Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays. *Mol. Cell. Proteomics* **13,** 1585–97 (2014).

30. Buus, S. *et al.* High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. *Mol. Cell. Proteomics* **11,** 1790–800 (2012).

31. Christiansen, A. *et al.* High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5,** 12913 (2015).

32. Valenta, R. *et al.* Recombinant allergens: what does the future hold? *J. Allergy Clin. Immunol.* **127,** 860–4 (2011).

33. Varshney, P. *et al.* A randomized controlled study of peanut oral immunotherapy: clinical desensitization and modulation of the allergic response. *J. Allergy Clin. Immunol.* **127,** 654–60 (2011).

34. Akdis, C. A. & Akdis, M. Mechanisms of allergen-specific immunotherapy and immune tolerance to allergens. *World Allergy Organ. J.* **8,** 17 (2015).

35. Schmid, J. M., Würtzen, P. A., Dahl, R. & Hoffmann, H. J. Pretreatment IgE sensitization patterns determine the molecular profile of the IgG4 response during updosing of subcutaneous immunotherapy with timothy grass pollen extract. *J. Allergy Clin. Immunol.* (2015). doi:10.1016/j.jaci.2015.05.023

36. Poulsen, L. K. & Hummelshoj, L. Triggers of IgE class switching and allergy development. *Ann. Med.* **39,** 440–56 (2007).

Discussion

In this dissertation, various studies have been presented. In Chapter 1, it was demonstrated that coupling HTS with phage display enabled the identification of epitopes in sera from peanut allergic patients. The same approach was applied in Chapter 2 where it was used to identify peptides that inhibited α-cobratoxin. HTS coupled with phage display was then used in an attempt to map epitopes of a therapeutic target using sera from immunized animals (Chapter 3). Finally, high-throughput peptide arrays were used to map peanut allergen epitopes and established individualized residue binding patterns and highly persistent epitopes (Chapter 4). This section will discuss the advantages and limitations of the applied techniques, with a focus on the coupling of HTS with phage display.

Deep sequencing helps identify rare phages

A main advantage of characterizing selected phage pools by deep sequencing is associated with the sequencing throughput. The high throughput allows an identification of the majority of the selected phages. This point was illustrated in the selection of α-cobratoxin binders where two phage clones were identified by traditional phage display, but the HTS revealed that the true number of unique phages were much larger. Analysis of the peptide data in Chapter 1 suggested that about 80-90% of the peptides identified by at least 2 independent sequencing reads were present in the previous round. This was similar to findings by others[1] and indicated that deep sequencing had identified a majority of the selected phages. Compared to IonTorrent sequencing (Chapter 1 & 2), the increase in sequencing output observed with the MiSeq platform in the α-cobratoxin study (Chapter 3) should enable a more exhaustive characterization of the phages, possibly approaching a full characterization of the phage pool.

The sequencing platforms appear to have the capacity to exhaustively characterize the entire phage population; however, it remains to be determined whether it is cost-effective and whether the rare phages contribute additional information. Supporting the idea that rare phages are important, they contributed to the cluster that identified the Ara h 1 epitope in Chapter 1 (see Appendix A, Suppl. Figure S5). However, this cluster would probably had been identified without these rare phages, since their rank values were very low. Nonetheless, the importance of the rare phage is supported by the

observation that the epitope would have been missed for several samples if it had not been for the rare peptides, as discussed in the study[2]. This example illustrates one of the potential advantages of HTS, specifically the identification of rare phages in a large population. Coupling phage display with HTS can improve the detection of relevant phages with low prevalence. This suggests that the deep sequencing can be used to identify peptides of interest in phage pools that are obtained from phage experiments that have not been fully optimized. The optimization can be a laborious part of phage display experiments and the lack of convincing differences to control biopannings can lead to a discontinuation of the experiments. Deep sequencing may enable the identification of relevant phages in a less-than-optimal setting.

How many selection rounds are required in phage display?

Another interesting aspect of the deep sequencing of phages is the potential to reduce the number of phage selection rounds. By performing fewer selection rounds the workload can be decreased and the selection bias can be reduced which should diminish the loss of relevant sequences[1,3,4]. The idea had previously been proposed, however, it has not been applied[1]. In Chapter 1 the principle was demonstrated, as a peptide motif clearly emerged in the second selection round. However, the identified epitope could probably have been identified in the first selection round, since there was a cluster with a borderline p-value (Chapter 1, Figure 4) and alignment of peptides (Chapter 1, Figure 5) also suggested the epitope location in the first selection round.

In the α-cobratoxin experiments the binders were identified by prevalence and one toxin binder entered the top 3 most frequent peptides in each selection round. Thus, all of the best binders could be identified by the third selection round, which was less than the five rounds applied in the traditional phage display approach. It has not been tested whether traditional clone picking and Sanger sequencing could have identified the binder after 3 selection rounds.

Nonetheless, it seems apparent that HTS insights into the phage pool can assist the identification of relevant phages using fewer selection rounds than would be required by traditional phage display. The required number of rounds likely depends on the experimental setup including selection scheme and selection pressure. This may need to be determined experimentally for each setup, but the number of required rounds might be similar for comparable setups, e.g. selection for peptide binders. After

determining an optimal number of selection rounds, it should not be necessary to sequence the output of all selection rounds, which was done in the present studies.

Can the throughput of phage display be increased?

Phage display is a relatively laborious process and it may be more widely applied and used for new applications if the throughput can be increased. By coupling phage display with HTS and reducing the number of selection rounds, the effort would slightly decrease. However, a more substantial increase in throughput would be obtained through skipping of the manual clone picking and functional tests of the clones. The functional testing is limiting the number of samples that can be analyzed in parallel. With HTS, the phage eluates are analyzed directly after PCR and thus an increased number of samples can be processed, e.g. as seen in Chapter 1 with biopanning against 12 samples in parallel. Finally, the more thorough characterization of the selected phage pools, discussed above, also represent an increase in throughput over conventional phage display.

A major issue concerning throughput is the increased amount of data analysis that comes with deep sequencing. The large data amount spurs a need for programming skills or dedicated collaboration with bioinformaticians. Besides the need for specialized skills, the process of data treatment, filtering and analysis is time-consuming. This is especially true when new bioinformatic pipelines have to be developed or a new approach is applied in the analysis, e.g. the clustering approach applied in Chapter 1. In the studies presented here, the bioinformatic pipelines were established from the beginning, which required a substantial investment of time. However, in subsequent projects such as the α-cobratoxin study (Chapter 2) or the therapeutic target study (Chapter 3) only minor adjustments were needed for the data treatment and, accordingly, this process was much faster. The lack of standardized analytical approaches as well as bioinformatic tools significantly slows the data processing. However, with the development of such standardized programs, the data analysis can become relatively fast. This is particularly the case if similar data sets are analyzed with the same objective e.g. identification of binding motifs or frequent peptides. Until standardized approaches are achieved the deep sequencing analysis will be more time-consuming than regular phage display, however, it will still be possible to analyze more samples in parallel and with a deeper insight into the selected phage pool.

Identification of relevant peptides

As has been pointed out in several chapters, HTS provides insights into the phage selection process and the convergence on fewer unique phages. The observations are expected, since phage display theory is based on the idea of affinity selection, where each biopanning round leads to an increase in specificity of the phage-library and a corresponding decrease in diversity[3,5].

How can one identify phage peptides of interest from a diverse phage population characterized by HTS? Overall, there seem to be four approaches: they can either be selected among a) the most frequent peptides or b) peptides that are highly enriched during selection or c) similar peptides that give rise to a selection motif or d) peptides that align to a known target. This identification process can be coupled with contrasting or filtering, where the peptide output is compared to known TUP sequences, library peptides or parallel control experiments.

Ryvkin *et al.* used frequency as a selection criteria, since they used the twenty most prevalent peptides to identify an epitope of a monoclonal antibody[6]. This was similar to the approach taken in the α-cobratoxin chapter, where prevalence appeared to be associated with increased binding affinity for the target. Prevalence appears to be a good indicator if the phage selection scheme is optimized and stringent and there is a low number of confounding sequences e.g. TUPs. It may be the best criteria for relatively simple phage selection schemes against a purified homogenous target where the selections are appropriately optimized.

Hoen *et al.* filtered out peptides that overlapped with known TUPs in the SAROTUP database[7] and library sequences and selected candidates of interest among the frequent peptides as well as peptides that were highly enriched compared to the library[1]. Enrichment was also a considered in other studies[8,9]. The enrichment approach seems to be a relevant parameter, however, in Chapter 1 we observed that TUP sequences also undergo massive enrichment[2]. Furthermore, this approach requires exhaustive sequencing of the original library or early selection rounds.

Selection of similar peptides that form a motif were shown to indicate true peptide binders in Chapter 1. In a recent study, Rebollo *et al.* used the same approach to identify target-binding motifs[10], which could be a promising application of this approach. However, the method relies on the selection of related peptides which was e.g. not observed in the α-cobratoxin study. This study employed a less diverse library, which could be part of the explanation. Overall, the motif selection approach seems promising, but it may not work for all applications.

Finally, studies by Liu *et al.* and Ryvkin *et al.* have identified the relevant peptides by alignment to a known target[6,11]. This is similar to the alignment approach applied in Figure 5 in Chapter 1. This approach was successful since there was a large overlap between peptides that aligned well to the primary sequence of Ara h 1 and peptides that clustered in patient-specific clusters. However, this approach is limited to known targets and linear epitopes. In addition, with a large number of peptides it is always possible to align many of them to the target so it is essential to have proper control experiments and control alignments for comparison (see e.g. Figure 4 in Chapter 3).

So far, there is no single approach that consistently seems to work for identification of relevant peptides. The proper approach depends on the context and, at present, the best strategy may be to investigate all four approaches. Future studies may help elucidate which conditions warrant one approach over another.

Handling TUPs experimentally and analytically

In order to get rid of irrelevant TUP peptides some studies have attempted at filtering out certain peptides before the remaining peptides were investigated for relevant candidates. As an example, Hoen *et al.* filtered out peptides that were in the SAROTUP database or were prevalent in the partial sequencing of the phage library[1]. Ryvkin *et al.* filtered out peptides that were identified in a parallel, but unrelated, phage selection[6]. In Chapter 2 we took a different approach. Instead of applying a filter that removed peptides, which were exact matches to peptides in control experiments, we kept all peptides and used them for clustering. The presence of control peptides were then weighed in, when the clusters were evaluated. This approach has the benefit that it is not limited to the removal of identical peptides, but also takes peptides that are similar to control peptides into account. It may be relevant to compare with the SAROTUP database[7], however, the database is based on few studies and contains relatively few motifs. When we compared the output from the peanut sera study with the SAROTUP database there was very limited overlap (data not shown). And there was no overlap when the peptides were compared to the 'parasitic' peptides identified by Mathocko *et al.*[4] using the same phage library. The experimental context and the batch of the phage library appear to affect which TUPs emerge. In addition, we observed an overlap between TUPs and prevalent library phages (Chapter 3). Based on these studies an optimal experimental setup should probably include parallel control selections that can be used for comparison combined with the sequencing of a portion of the phage library.

The issue with TUPs that arise during amplification can potentially also be addressed experimentally. First of all, if the phages can be amplified sufficiently by PCR without involving an amplification step in bacteria, that would decrease the proliferation bias. We have attempted direct sequencing of the phage eluate and the procedure was somewhat successful (data not shown). Drawbacks included an increased number of required PCR cycles, and the use of the entire eluate so that nothing remained for additional selection rounds. Most importantly, though, the reaction did not work reliably. Further optimization, possibly incorporating new kits from Qiagen for phage DNA purification (the QIAprep Spin M13 Kit), may increase the reliability of this approach. Another approach to reduce the bias is to limit the competition between phages during amplification by performing the proliferation step in emulsion droplets[4,12]. However, this approach requires some technical experience with e.g. microfluidic systems and emulsion droplets.

Low-frequency derivative peptides in deep sequencing

In Chapter 1, we became aware of the issue with derivate peptides. Highly abundant sequences and HTS sequencing errors gave rise to a population of peptides that differed in one or two bases from the abundant sequence and consequently appeared as a group of sequences with high similarity to the abundant 'parent' sequence. This phenomenon is not observed in regular genomic HTS as the same sequences are sampled fewer times (e.g. to 20-30x coverage). After aligning e.g. 30 sequences to the same area, deviations from the majority of alignments are discarded as HTS errors. In the analysis of phage pools it become a critical issue due to the abundance of some sequences; especially if the ambition is to use clustering of similar sequences to identify relevant peptides.

We empirically determined how many derivative peptides were observed and then set up a simple threshold to eliminate the suspected peptides. The threshold was based on how many times more prevalent the 'parent' sequence was. A similar approach was recently taken in other studies[4,10], though Rebollo *et al.* altered the derivative peptides back to match the abundant 'parent' sequence. Since there may be differences between experimental setups and HTS platforms, the empirical approach may currently be the best. However, the phenomenon could clearly benefit from dedicated studies in order to elucidate general error-correction algorithms[13]. This has been carried out for viral sequencing data and been used to establish models that may assist the identification of HTS errors[14].

Characterization of conformational epitopes

Most of the epitopes identified in food allergy, including those that have been characterized in microarray studies, have been linear epitopes[15]. However, studies have indicated that conformational epitopes are also important in food allergy[16–24]. For example, in studies of patients with milk and egg allergy, it has been described how those who generally reacted to linear IgE epitopes were more likely to have persistent allergy than those who reacted to conformational epitopes[25–28]. Phage display has the potential to identify conformational epitopes, since the libraries often have a very high diversity. Ryvkin *et al.* succeeded in mapping the conformational epitope of a monoclonal antibody by HTS-coupled phage display[6], thus demonstrating the principle. The conformational epitopes are difficult to map, since the selected phage peptides bears little resemblance to the target sequence. Instead, the peptides are mimotopes that emulate the target protein structure. Consequently, the peptides have to be compared to the surface of the target structure using bioinformatic tools[29]. However, these tools can often align peptides to any target so care has to be taken in the interpretation and appropriate controls need to be included. Ideally, the prediction tools should be improved to solve this challenging task.

While the clustering tool employed in Chapter 1 could potentially have identified conformational mimotope clusters, a linear epitope cluster was the only one to emerge. Potential explanations include that the patients only have linear epitopes or that any mimotopes are so diverse that they do not align and cluster well. Furthermore, the 7-mer peptide applied in this study may not be able to adequately mimic a conformational epitope and phage libraries with e.g. 12-mer peptides or structurally restricting cysteine bridges may be better suited for epitope identification.

Next-generation peptide microarrays for epitope mapping

As demonstrated in Chapter 1 & 4, peptide microarrays are highly suited for epitope mapping, especially linear epitope mapping. The next-generation peptide arrays available from Roche NimbleGen have a very high throughput, e.g. the 2.1 million peptides available on a slide have recently been used to cover the entire human proteome[30]. The arrays can be tailored to provide great insights into epitope binding patterns for selected proteins as illustrated in Chapter 4. However, when the entire slide is used for a single sample it is relatively expensive, thus splitting the array into 12 wells, as were done in our peptide array study, can be more cost-effective.

The peptide arrays identified the epitope that was also identified by phage display in Chapter 1. So, what is the advantage of the phage display approach that warrants the extra labor that the technique requires? The main answer relates to phage diversity. While the peptide array can potentially test 2.1 million peptides, it is expensive (compared to conventional phage display) and the diversity is a lot less than the reported diversity of 2×10^9 for the Ph.D.7™ library. Consequently, the arrays have to be designed and produced with a specific purpose in mind, while the same phage library can be used for many applications. The diversity also eases the identification of peptide clusters and conformational epitopes, as described above.

From a technical perspective, we observed a very high signal-to-noise ratio with the new microarrays in Chapter 4. The sensitivity increased from the observations made in Chapter 1 with new and improved microarrays used in Chapter 4. This lead to an increase in the number of detected epitopes. E.g. whereas a single Ara h 1 epitope at position 136 was observed in the initial studies, later studies revealed that there were several additional epitopes for P6 in Chapter 4. These findings indicate that the original phage experiments missed some epitopes. It also indicate that there was something distinctive about the 136 epitope, since it arose from the phage experiments as well as the microarray experiments. The explanation could be linked to antibody concentration or affinity for the antibodies that recognize this particular epitope.

Concluding remarks

The coupling of phage display with HTS holds some potential advantages over conventional phage display. These include the identification of rare phages in a large population, the skipping of laborious steps and the use of fewer selection rounds. The approach also provides an improved understanding of the phage selection process. However, there are several areas that need further studies to more fully determine the potential of the approach. In particular, there is a need to establish general bioinformatic tools and ways to identify the relevant phages. Nonetheless, the increased throughput combined with the in-depth data output holds the promise to expand the use of phage display to new applications, including the analysis of complex biological samples.

References

1. 'T Hoen, P. a C. *et al.* Phage display screening without repetitious selection rounds. *Anal. Biochem.* **421,** 622–631 (2012).

2. Christiansen, A. *et al.* High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Sci. Rep.* **5,** 12913 (2015).

3. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16,** 1776–1803 (2011).

4. Matochko, W. L., Cory Li, S., Tang, S. K. Y. & Derda, R. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res.* **42,** 1784–98 (2014).

5. Smith, G. P. & Petrenko, V. A. Phage Display. *Chem. Rev.* **97,** 391–410 (1997).

6. Ryvkin, A. *et al.* Deep panning: Steps towards probing the IgOme. *PLoS One* **7,** e41469 (2012).

7. Huang, J., Ru, B., Li, S., Lin, H. & Guo, F. B. SAROTUP: Scanner and reporter of target-unrelated peptides. *J. Biomed. Biotechnol.* **2010,** 101932 (2010).

8. Ngubane, N. a C. *et al.* High-throughput sequencing enhanced phage display identifies peptides that bind mycobacteria. *PLoS One* **8,** e77844 (2013).

9. Hu, D. *et al.* Effective Optimization of Antibody Affinity by Phage Display Integrated with High-Throughput DNA Synthesis and Sequencing Technologies. *PLoS One* **10,** e0129125 (2015).

10. Rentero Rebollo, I., Sabisz, M., Baeriswyl, V. & Heinis, C. Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides. *Nucleic Acids Res.* **42,** e169 (2014).

11. Liu, X. *et al.* Serum Antibody Repertoire Profiling Using In Silico Antigen Screen. *PLoS One* **8,** e67181 (2013).

12. Matochko, W. L. *et al.* Uniform amplification of phage display libraries in monodisperse emulsions. *Methods* **58,** 18–27 (2012).

13. Matochko, W. L. & Derda, R. Error analysis of deep sequencing of phage libraries: Peptides censored in sequencing. *Comput. Math. Methods Med.* **2013,** (2013).

14. Orton, R. J. *et al.* Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* **16,** 229 (2015).

15. Bøgh, K. L. *et al.* IgE versus IgG4 epitopes of the peanut allergen Ara h 1 in patients with severe allergy. *Mol. Immunol.* **58,** 169–176 (2013).

16. Pomés, A. Relevant B cell epitopes in allergic disease. *Int. Arch. Allergy Immunol.* **152,** 1–11 (2010).

17. Van Regenmortel, M. H. V. What is a B-cell epitope? *Methods Mol. Biol.* **524,** 3–20 (2009).

18. Clement, G. *et al.* Epitopic characterization of native bovine beta-lactoglobulin. *J. Immunol. Methods* **266,** 67–78 (2002).

19. San Miguel-Moncín, M. *et al.* Lettuce anaphylaxis: identification of a lipid transfer protein as the major allergen. *Allergy* **58,** 511–7 (2003).

20. Sathe, S. K., Teuber, S. S. & Roux, K. H. Effects of food processing on the stability of food allergens. *Biotechnol. Adv.* **23,** 423–9 (2005).

21. Denery-Papini, S. *et al.* Immunoglobulin-E-binding epitopes of wheat allergens in patients with food allergy to wheat and in mice experimentally sensitized to wheat proteins. *Clin. Exp. Allergy* **41,** 1478–92 (2011).

22. Albrecht, M. *et al.* Relevance of IgE binding to short peptides for the allergenic activity of food allergens. *J. Allergy Clin. Immunol.* **124,** 328–36, 336.e1–6 (2009).

23. Starkl, P. *et al.* An unfolded variant of the major peanut allergen Ara h 2 with decreased anaphylactic potential. *Clin. Exp. Allergy* **42,** 1801–12 (2012).

24. Apostolovic, D. *et al.* Reduction and alkylation of peanut allergen isoforms Ara h 2 and Ara h 6; characterization of intermediate- and end products. *Biochim. Biophys. Acta* **1834,** 2832–42 (2013).

25. Järvinen, K.-M. *et al.* Specificity of IgE antibodies to sequential epitopes of hen's egg ovomucoid as a marker for persistence of egg allergy. *Allergy* **62,** 758–765 (2007).

26. Vila, L. *et al.* Role of conformational and linear epitopes in the achievement of tolerance in cow's milk allergy. *Clin. Exp. Allergy* **31,** 1599–606 (2001).

27. Järvinen, K. M., Chatchatee, P., Bardina, L., Beyer, K. & Sampson, H. A. IgE and IgG binding epitopes on alpha-lactalbumin and beta-lactoglobulin in cow's milk allergy. *Int. Arch. Allergy Immunol.* **126,** 111–8 (2001).

28. Cooke, S. K. & Sampson, H. A. Allergenic properties of ovomucoid in man. *J. Immunol.* **159,** 2026–32 (1997).

29. Bublil, E. M. *et al.* Stepwise prediction of conformational discontinuous B-cell epitopes using the mapitope algorithm. *Proteins Struct. Funct. Genet.* **68,** 294–304 (2007).

30. Forsström, B. *et al.* Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays. *Mol. Cell. Proteomics* **13,** 1585–97 (2014).

Appendix A

Supplementary information – Chapter 1

Supplementary information - "High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum"

Anders Christiansen[1], Jens V. Kringelum[2], Christian S. Hansen[2], Katrine L. Bøgh[3], Eric Sullivan[4], Jigar Patel[4], Neil M. Rigby[5], Thomas Eiwegger[6], Zsolt Szépfalusi[6], Federico de Masi[2], Morten Nielsen[2,7], Ole Lund[2], Martin Dufva[1]*

[1] Department of Micro- and Nanotechnology, Technical University of Denmark, Kgs. Lyngby, Denmark,

[2] Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark

[3] National Food Institute, Technical University of Denmark, Søborg, Denmark

[4] Roche NimbleGen, Madison, Wisconsin, the United States of America

[5] Institute of Food Research, Norwich, United Kingdom

[6] Department of Paediatrics, Medical University of Vienna, Vienna, Austria

[7] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

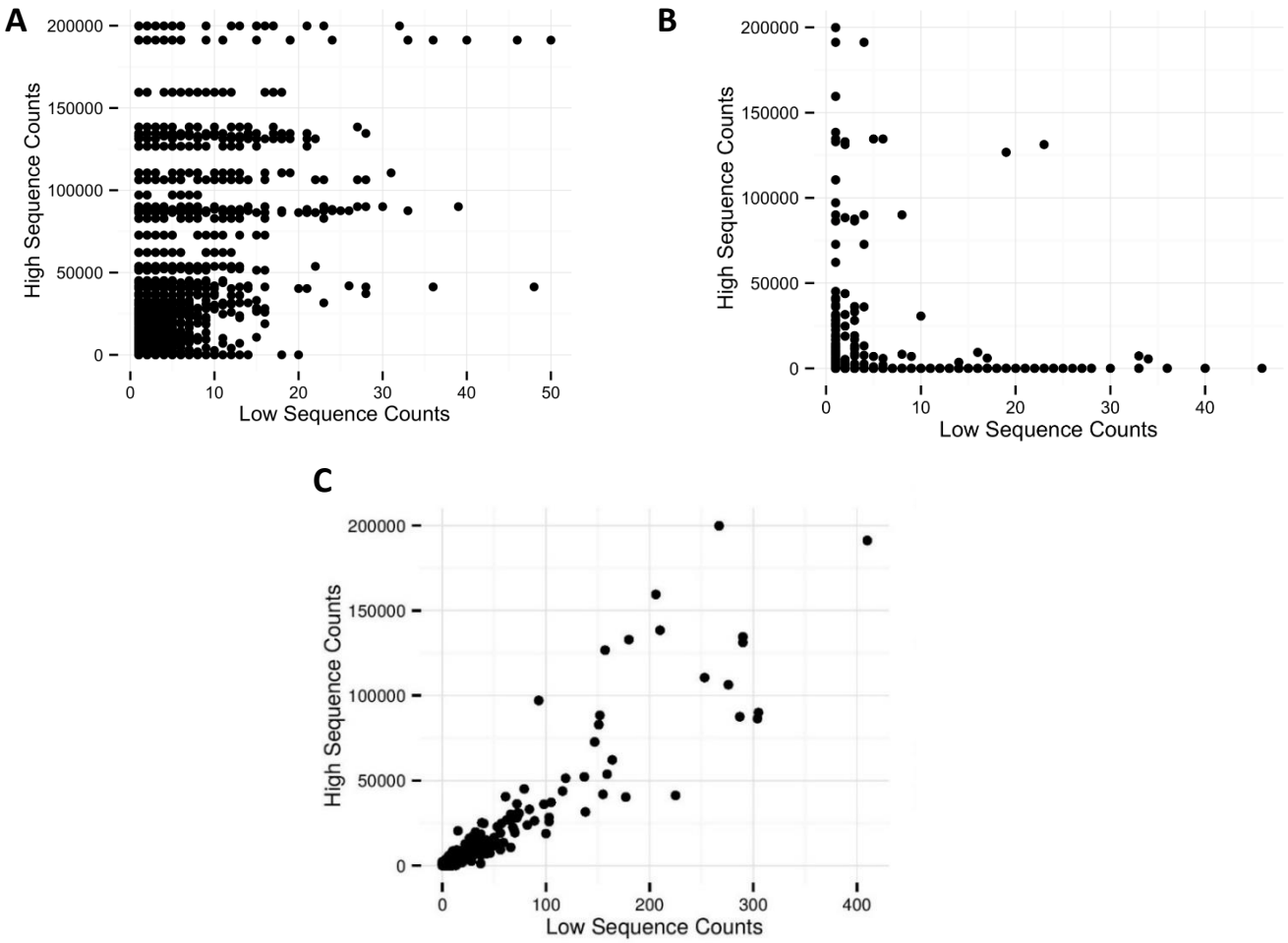
* Correspondence and requests for materials should be addressed to M.D. (martin.dufva@nanotech.dtu.dk)

5´-*AAGAGGATTC* GAT TTC GCA ATT CCT TTA GTG
5´-*CTGCAAGTTC* GAT TTC GCA ATT CCT TTA GTG
5´-*TCTAACGGAC* GAT TTC GCA ATT CCT TTA GTG
5´-*TTGGAGTGTC* GAT TTC GCA ATT CCT TTA GTG
5´-*TCTGGATGAC* GAT TTC GCA ATT CCT TTA GTG
5´-*TCTATTCGTC* GAT TTC GCA ATT CCT TTA GTG
Reverse: 5´-GCT AAA CAA CTT TCA ACA G

Supplementary Table S1. Primer and barcode DNA sequences used for PCR. The barcode element has been italicised. The reverse primer in the bottom row was used for all reactions.

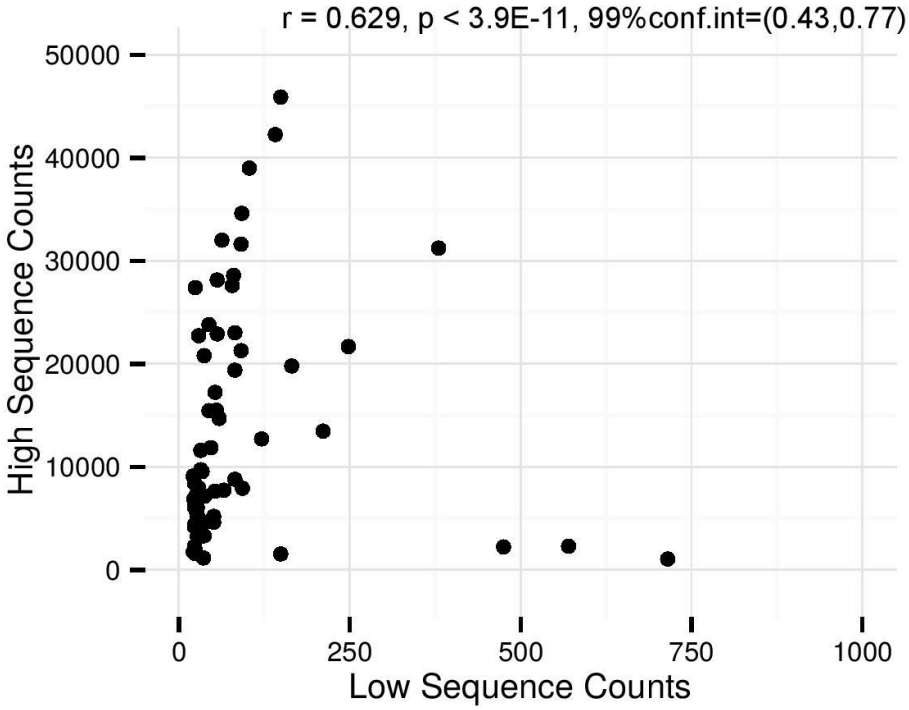
Sequencing Run	Patient	Selection round	#reads w barcode	Peptides called	NNK filtering	Quality filtering
1	C2P	1	270337	199212 (73%)	198283 (73%)	151466 (56%)
1	C2P	2	348124	179438 (51%)	177502 (50%)	93727 (26%)
1	C2P	3	341409	152896 (44%)	150871 (44%)	65241 (19%)
1	P4_12	1	337208	193715 (57%)	191271 (56%)	123682 (36%)
1	P4_12	2	267119	161249 (60%)	160271 (59%)	110706 (41%)
1	P4_12	3	317513	216935 (68%)	215009 (67%)	141279 (44%)
2	C1	1	372884	320282 (85%)	318595 (85%)	209704 (56%)
2	C1	2	247717	201535 (81%)	200532 (80%)	141377 (57%)
2	C1	3	229598	188859 (82%)	188057 (81%)	110312 (48%)
2	C3	1	493126	433610 (87%)	429495 (87%)	312966 (63%)
2	C3	2	315301	275409 (87%)	274312 (87%)	218529 (69%)
2	C3	3	444588	398677 (89%)	397525 (89%)	352758 (79%)
3	C2	1	187617	145280 (77%)	144427 (76%)	87672 (46%)
3	C2	2	163062	139999 (85%)	139370 (85%)	92461 (56%)
3	C2	3	191137	165666 (86%)	164854 (86%)	110574 (57%)
3	C4	1	314801	264167 (83%)	260828 (82%)	181312 (57%)
3	C4	2	150530	111450 (74%)	110413 (73%)	73886 (49%)
3	C4	3	211778	160403 (75%)	158561 (74%)	96523 (45%)
4	P2	1	389423	336846 (86%)	334691 (85%)	225844 (57%)
4	P2	2	248397	222564 (89%)	221625 (89%)	178082 (71%)
4	P2	3	269238	239172 (88%)	238660 (88%)	198232 (73%)
4	P4	1	470407	373218 (79%)	370282 (78%)	247124 (52%)
4	P4	2	277800	223134 (80%)	222006 (79%)	149424 (53%)
4	P4	3	460451	395964 (85%)	392917 (85%)	278072 (60%)
5	P1_12	1	259866	148832 (57%)	146965 (56%)	102337 (39%)
5	P1_12	2	255146	150404 (58%)	148385 (58%)	111000 (43%)
5	P1_12	3	247601	135410 (54%)	134204 (54%)	99317 (40%)
5	C3P	1	301990	173918 (57%)	172066 (56%)	111241 (36%)
5	C3P	2	234836	103078 (43%)	102275 (43%)	57059 (24%)
5	C3P	3	246973	99754 (40%)	98743 (39%)	70327 (28%)
6	P3	1	573487	459349 (80%)	456976 (79%)	304519 (53%)
6	P3	2	277907	201369 (72%)	200826 (72%)	126143 (45%)
6	P3	3	168015	94123 (56%)	93657 (55%)	65886 (39%)
6	P1	1	499211	425951 (85%)	424085 (84%)	307978 (61%)
6	P1	2	349713	317818 (90%)	316771 (90%)	256409 (73%)
6	P1	3	362151	329429 (90%)	328668 (90%)	262470 (72%)

Supplementary Table S2. Number of sequences in each step of the processing. The provided numbers are the number of sequences that remains at the particular stage. The percentage denotes the percentage of remaining sequences compared to the initial assignment based on barcodes. In brief the steps were (i) associating sequences with a sample based on barcode, (ii) calling 7-mer peptides and (iii) filtering peptides based NNK codon pattern and (iv) quality filtering. The steps are covered in greater detail in the Materials and Methods section.



Supplementary Figure S1. (A) Derivative sequences due to high-throughput sequencing errors. The relationship between the prevalence of f a high-frequent peptide ("High Sequence Counts") and the prevalence of similar peptides where the DNA sequence varies at a single position ("Low Sequence Counts") has been plotted. Note that the axes have different scales. The horizontal rows of observations that are visible at high prevalences corresponds to the various derivative sequences that are observed from a single high-frequent peptide. **(B)** As (A) but plotting the prevalence of derivative sequences where there are two deviations in the DNA sequences compared to the high-frequent peptide. **(C)** As (A) but depicting the sum of all derivative sequences on the "Low Sequence Counts" axis for each high-frequent peptide. A linear relationship between the prevalence of the high-frequent peptide and the sum of the derivative sequences is observed, where peptides with higher prevalences give rise to more derivative sequences. Such a relation is to be expected if the derivative sequences are due to sequencing errors.

Threshold details: The applied thresholds of 500 for single-base deviations (shown in A) and 10,000 for double-base deviations (shown in B) where empically determined. Taking the 500-fold threshold as an example, sequences were removed if they were at least 500 times less prevalent than a similar sequence (differing by a single base pair) in the same sample. Translating this to a per-base error rate was done by the formula $P^{20} \times ((1-P) / 3) = 1/500 \times P\text{^}21$, where P is the per-base accuracy and 1-P is the per-base error rate. (1-P) /3 assumes that the change to any of the three other bases are equally distributed. Solving for P yields P = 0.994, thus the 500-fold threshold corresponds to a per-base sequencing error rate of 0.6%.



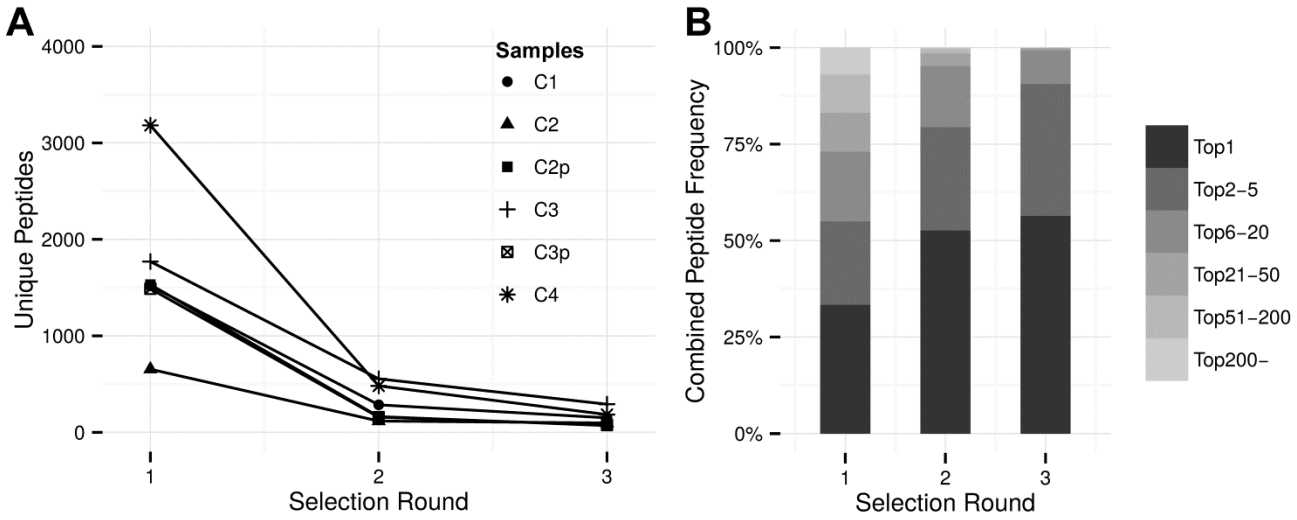
Supplementary Figure S2. Derivative sequences due to high-throughput sequencing errors give rise to "spill-over" between samples. The relationship between the prevalence of a high-frequent peptide ("Count High") and the prevalence of the same peptides in another subject ("Count Low"), analysed on the same sequencing chip has been plotted. Note that the axes have different scales. A maximum of 50,000 and 1,000 is plotted for "Count High" and "Count Low", respectively, has been plotted. The summed prevalence for each peptide across the three selection rounds is depicted. A Pearson correlation coefficient has been calculated and included.

Sequencing Run	Patient	Selection round	Peptide sequences	"Derivative" filtering	"Spill-over" filtering
1	C2P	1	151466	151025 (99%)	150936 (99%)
1	C2P	2	93727	93495 (99%)	93489 (99%)
1	C2P	3	65241	65010 (99%)	64979 (99%)
1	P4_12	1	123682	123432 (99%)	123378 (99%)
1	P4_12	2	110706	110385 (99%)	110380 (99%)
1	P4_12	3	141279	140905 (99%)	140880 (99%)
2	C1	1	209704	209320 (99%)	208751 (99%)
2	C1	2	141377	141003 (99%)	140477 (99%)
2	C1	3	110312	110128 (99%)	109509 (99%)
2	C3	1	312966	312506 (99%)	311830 (99%)
2	C3	2	218529	218051 (99%)	217719 (99%)
2	C3	3	352758	352267 (99%)	351732 (99%)
3	C2	1	87672	87501 (99%)	87321 (99%)
3	C2	2	92461	92260 (99%)	92071 (99%)
3	C2	3	110574	110383 (99%)	110095 (99%)
3	C4	1	181312	181039 (99%)	180491 (99%)
3	C4	2	73886	73665 (99%)	73481 (99%)
3	C4	3	96523	96349 (99%)	96013 (99%)
4	P2	1	225844	225377 (99%)	224941 (99%)
4	P2	2	178082	177633 (99%)	177381 (99%)
4	P2	3	198232	197940 (99%)	197718 (99%)
4	P4	1	247124	246787 (99%)	245951 (99%)
4	P4	2	149424	148951 (99%)	148441 (99%)
4	P4	3	278072	277535 (99%)	276623 (99%)
5	P1_12	1	102337	102163 (99%)	102149 (99%)
5	P1_12	2	111000	110699 (99%)	110692 (99%)
5	P1_12	3	99317	98988 (99%)	98969 (99%)
5	C3P	1	111241	110945 (99%)	110894 (99%)
5	C3P	2	57059	56864 (99%)	56836 (99%)
5	C3P	3	70327	70080 (99%)	70048 (99%)
6	P3	1	304519	303877 (99%)	302501 (99%)
6	P3	2	126143	125726 (99%)	124944 (99%)
6	P3	3	65886	65616 (99%)	63626 (96%)
6	P1	1	307978	307130 (99%)	306615 (99%)
6	P1	2	256409	255457 (99%)	255134 (99%)
6	P1	3	262470	261885 (99%)	261614 (99%)

Supplementary Table S3. Number of sequences filtered due to high-throughput sequencing errors. "Derivative" filtering comprise the removal of the derivative sequences plotted in Supplementary Figure S1, i.e. peptides in a sample that are very similar to a high-frequent peptide in the same sample. "Spill-over" filtering encompasses the removal of the derivative sequences plotted in Supplementary Figure S2, i.e. peptides in a sample that are very similar to a high-frequent peptide in a different subject, analysed on the same sequencing chip. The provided numbers are the number of peptide sequences that remains at the particular stage. The percentage denotes the percentage of remaining peptides after the specific filtering step.

Sequencing Run	Patient	Selection round	Unique Peptides	"Derivative" filtering	"Spill-over" filtering
1	C2P	**1**	1655	1552 (93%)	1531 (92%)
1	C2P	2	262	166 (63%)	163 (62%)
1	C2P	3	147	81 (55%)	70 (47%)
1	P4_12	1	2181	2033 (93%)	2023 (92%)
1	P4_12	2	362	242 (66%)	239 (66%)
1	P4_12	3	206	119 (57%)	110 (53%)
2	C1	1	1839	1582 (86%)	1519 (82%)
2	C1	2	456	358 (78%)	285 (62%)
2	C1	3	250	208 (83%)	150 (60%)
2	C3	1	2045	1846 (90%)	1770 (86%)
2	C3	2	811	625 (77%)	555 (68%)
2	C3	3	474	365 (77%)	291 (61%)
3	C2	1	799	696 (87%)	654 (81%)
3	C2	2	228	165 (72%)	117 (51%)
3	C2	3	203	156 (76%)	97 (47%)
3	C4	1	3367	3210 (95%)	3182 (94%)
3	C4	2	637	503 (78%)	481 (75%)
3	C4	3	307	205 (66%)	184 (59%)
4	P2	1	1660	1529 (92%)	1429 (86%)
4	P2	2	349	227 (65%)	168 (48%)
4	P2	3	218	145 (66%)	89 (40%)
4	P4	1	2484	2279 (91%)	2233 (89%)
4	P4	2	614	459 (74%)	437 (71%)
4	P4	3	419	293 (69%)	253 (60%)
5	P1_12	1	1974	1875 (94%)	1869 (94%)
5	P1_12	2	436	292 (66%)	286 (65%)
5	P1_12	3	203	88 (43%)	82 (40%)
5	C3P	1	1681	1504 (89%)	1488 (88%)
5	C3P	2	260	167 (64%)	157 (60%)
5	C3P	3	186	88 (47%)	76 (40%)
6	P3	1	3078	2711 (88%)	2611 (84%)
6	P3	2	643	550 (85%)	469 (72%)
6	P3	3	594	540 (90%)	436 (73%)
6	P1	1	2255	1771 (78%)	1622 (71%)
6	P1	2	754	522 (69%)	410 (54%)
6	P1	3	390	278 (71%)	197 (50%)

Supplementary Table S4. Number of unique peptide sequences filtered due to high-throughput sequencing errors. The table is similar to Supplementary Table S2, however, the number of unique peptides is given instead of total peptide sequence numbers. "Derivative" filtering comprise the removal of the derivative sequences plotted in Supplementary Figure S1, i.e. peptides in a sample that are very similar to a high-frequent peptide in the same sample. "Spill-over" filtering encompasses the removal of the derivative sequences plotted in Supplementary Figure S2, i.e. peptides in a sample that are very similar to a high-frequent peptide in a different subject, analysed on the same sequencing chip. The percentage denotes the percentage of remaining unique peptides after the specific filtering step.



Supplementary Figure S3. The selection process in control samples. **(A)** A plot of the number of unique peptides identified in each control sample for selection round 1-3. **(B)** A stacked bar chart showing the combined frequency of certain rank intervals. Specifically, the frequency of the most frequent peptide (Top1) along with the combined frequencies of the peptides ranked 2-5, 6-20, 21-50, 51-200 and below 200, based on their frequency in a sample. The average of the combined frequencies for all control samples is shown.

PATIENT: P1					
WSLSELH	14.3%	VSRDTPQ	33.9%	VSRDTPQ	73.1%
DGRYYIN	9.1%	DGRYYIN	20.2%	DGRYYIN	9.5%
VSRDTPQ	4.5%	WSLSELH	15.8%	WSLSELH	8.7%
MHTVAVQ	4.2%	HWRLPLH	7.4%	HLNQQNH	3.4%
HLNQQNH	3.5%	HLNQQNH	4.6%	HWRLPLH	0.9%
APRNVPP	2.4%	IWRLPTH	4.2%	IWRLPTH	0.7%
WPFHGDN	1.5%	MHTVAVQ	2.4%	SYTDLLR	0.7%
QISAASQ	1.5%	VMPLDDV	1.4%	VMPLDDV	0.5%
LPYDHLP	1.5%	SYTDLLR	1.2%	YPWFIRA	0.5%
YPWFIRA	1.4%	YPWFIRA	0.9%	MHTVAVQ	0.3%

PATIENT: P2					
DNYDSTA	59.8%	DNYDSTA	74.0%	DNYDSTA	64.1%
MLHELLE	3.0%	VSGLLVE	7.3%	ENHVHVR	15.5%
RVASLAP	2.7%	ENHVHVR	5.3%	VSGLLVE	12.7%
ENHVHVR	2.6%	MLHELLE	4.3%	YPWFIRA	4.7%
NYFKHEA	2.3%	NSPELTS	2.1%	MLHELLE	2.5%
DAIPTSV	1.8%	DAIPTSV	1.7%	SLLGQTP	0.1%
VSGLLVE	1.8%	SLLGQTP	1.1%	NSPELTS	0.1%
ELRNENT	1.8%	YPWFIRA	0.7%	SPTTTYD	0.1%
NSPELTS	1.1%	NYFKHEA	0.5%	TALRTIS	0.1%
NGHSIHT	1.0%	ISPLSVP	0.5%	ISPLSVP	0.0%

PATIENT: P3					
EVMGRLA	10.4%	EVMGRLA	72.1%	EVMGRLA	64.9%
STTPHSR	7.5%	MVPPSYD	5.8%	MVPPSYD	8.7%
MVPPSYD	3.4%	QLIHWRH	4.8%	VKFDKYV	8.5%
SSIGGQD	2.9%	QLYREFN	3.6%	AMSHSKQ	5.4%
QLYREFN	2.8%	SSNQFHQ	1.5%	QLIHWRH	3.1%
YLGFDVH	1.8%	HSILSNL	1.5%	YLGFDVH	1.2%
SQNTKYI	1.5%	YLGFDVH	1.4%	WSLSELH	1.1%
QLIHWRH	1.3%	AQYVAVG	0.9%	HSILSNL	0.9%
MVQHYAE	1.2%	SSIGGQD	0.8%	LKFDTPI	0.9%
HSILSNL	1.2%	ATDEIVP	0.8%	QLYREFN	0.7%

PATIENT: P4					
SLLGQTP	16.5%	SLLGQTP	41.9%	SLLGQTP	50.0%
STIVAEM	6.8%	SAAWNKS	16.0%	SAAWNKS	18.9%
GFTFQPS	5.3%	STIVAEM	12.9%	STIVAEM	13.1%
SAAWNKS	4.7%	GTGSQAS	6.6%	GTGSQAS	6.2%
TNLSHVP	3.4%	GFTFQPS	5.9%	QTWLEMG	4.4%
QTWLEMG	3.3%	QTWLEMG	4.9%	GFTFQPS	2.5%
LLADMHA	2.6%	TNLSHVP	2.3%	KGIYHQV	0.8%
GTGSQAS	2.0%	SIIARPG	1.8%	STPATLI	0.7%
LVYVDMH	1.0%	STPATLI	1.2%	SLSSAWW	0.6%
ISYLSVT	0.9%	LLADMHA	0.8%	TNLSHVP	0.5%

PATIENT: P1_12					
AMSPLNK	16.2%	MPTWLHH	29.9%	GASESYL	36.4%
MPTWLHH	13.8%	AMSPLNK	17.5%	MPTWLHH	26.1%
NVRLPYQ	9.2%	GASESYL	10.4%	IENRIYR	13.4%
NLLDSLH	6.0%	GVFISYN	5.3%	AMSPLNK	8.6%
NPTHPIY	3.5%	FASRSDT	4.8%	FASRSDT	3.6%
FASRSDT	2.7%	IENRIYR	4.7%	NPPWFHT	2.2%
IENRIYR	2.3%	NVRLPYQ	3.5%	GVFISYN	2.2%
GVFISYN	2.2%	IPNGHFT	3.4%	IPNGHFT	1.4%
LSKINSP	2.1%	NLLDSLH	3.3%	NVRLPYQ	0.7%
GNEVMTY	1.6%	NPTHPIY	2.4%	NLLDSLH	0.7%

PATIENT: P4_12					
LSANHWV	24.5%	VPNIVTQ	48.7%	VPNIVTQ	75.5%
VPNIVTQ	7.5%	LSANHWV	24.3%	LSANHWV	5.6%
VTRDSNH	7.2%	SFNLPNT	5.1%	HGGVRLY	5.0%
SHPLWNS	5.6%	GMMSSPP	4.7%	SFNLPNT	3.4%
SFNLPNT	4.1%	YVTRTPY	1.8%	GMMSSPP	3.1%
LPFINSD	2.5%	HGGVRLY	1.3%	YVTRTPY	1.5%
HTNSAYI	1.7%	HYIDFRW	1.2%	VSYGVPM	0.8%
GMMSSPP	1.6%	ATKHYTT	1.0%	HYIDFRW	0.7%
DAIPTSV	1.5%	AVDPQYG	1.0%	KGIYHQV	0.3%
HAGFVPS	1.4%	DAIPTSV	0.8%	AVDPQYG	0.3%

CONTROL: C1					
EFARNSI	21.6%	EFARNSI	78.7%	EFARNSI	75.7%
LDIVDAP	3.6%	QHDTVPP	6.7%	HYIDFRW	18.7%
ALPLLDA	2.6%	HSTQLPY	3.5%	QHDTVPP	1.6%
FSTNFGN	2.5%	HYIDFRW	2.2%	FVPHKWY	1.1%
QHDTVPP	2.1%	FVPHKWY	1.9%	YGNSGIV	1.0%
HISGRPL	2.0%	AMSHSKQ	1.2%	AMSHSKQ	0.8%
VNGHYTI	2.0%	FCPDCKP	0.8%	FCPDCKP	0.4%
QIGHDGN	1.8%	TGKMLAD	0.6%	HSTQLPY	0.1%
GQAMNHT	1.8%	GAAYIRA	0.5%	NVAHKMF	0.1%
DFHPDVL	1.7%	NVAHKMF	0.5%	TGKMLAD	0.0%

CONTROL: C2					
VSRDTPQ	28.4%	VSRDTPQ	79.0%	VSRDTPQ	80.3%
SYTPGGH	16.7%	SAEKLLR	6.9%	SAEKLLR	13.2%
LIQPITT	7.4%	YNISVNK	3.0%	GQSEKHL	3.5%
QLYREFN	4.2%	QLYREFN	2.4%	YNISVNK	1.1%
MDTPDRI	3.5%	WVTDSSW	1.7%	WVTDSSW	0.8%
WVTDSSW	3.2%	HSPNSIK	1.4%	QLYREFN	0.4%
HSPNSIK	2.9%	GQSEKHL	1.3%	LPGNRLL	0.3%
RFDWSYP	2.9%	RFDWSYP	1.2%	RFDWSYP	0.3%
AYASDSY	2.7%	LPGNRLL	0.9%	HSPNSIK	0.0%
YNISVNK	2.6%	RTWEPYT	0.5%	RTWEPYT	0.0%

CONTROL: C3					
SNYHPSI	51.2%	SNYHPSI	61.0%	AERYPDS	56.8%
TTQVLEA	3.1%	KMISATE	6.2%	SNYHPSI	27.6%
KMISATE	2.9%	AERYPDS	5.2%	KMISATE	8.0%
LPQSWAM	1.8%	KMESGTA	3.7%	ELGTTQT	2.3%
ATAEWHP	1.7%	AKYEAGS	3.2%	TTQVLEA	1.0%
APANTFT	1.7%	TTQVLEA	1.9%	AKYEAGS	0.7%
VPPTEHP	1.3%	TEQKIEA	1.9%	GWETRME	0.5%
LGQLFPQ	1.3%	TITAMQG	1.4%	NDRPHMP	0.3%
MIHPRDQ	1.1%	AERILSV	1.3%	ADLMHFT	0.3%
FTQPLYK	1.1%	GWETRME	1.2%	VPPTEHP	0.3%

CONTROL: C4					
KSLNSTI	17.2%	VPILNSL	19.8%	WSLSELH	20.6%
VPILNSL	7.1%	WSLSELH	17.0%	HAAWHFI	19.3%
SMNETFA	6.6%	ENHVHVR	13.2%	ENHVHVR	16.7%
WSLSELH	3.7%	QETRGTY	4.8%	VPILNSL	7.6%
NTDTTRE	3.3%	IDNSHTH	4.0%	VTLPPSS	3.8%
VNQSWDL	2.0%	TTANVRI	3.9%	TTANVRI	3.4%
RDEVYFH	1.9%	VNQSWDL	3.1%	ESKWMPL	3.2%
NEAPRHA	1.7%	HAAWHFI	2.9%	IDNSHTH	3.2%
IDRTQFM	1.5%	EPAIATP	2.3%	QETRGTY	3.0%
NTVGTVQ	1.3%	NEAPRHA	2.3%	GTMHSMK	2.6%

CONTROL: C2p					
MGQPTVK	58.0%	QQLAVDT	39.8%	QQLAVDT	64.5%
GPLKTWK	18.6%	MGQPTVK	16.2%	ESRVMSR	14.4%
GTLASLN	5.4%	TTQVLEA	8.8%	WEGPQPK	7.9%
KVKTYDM	3.4%	WEGPQPK	6.9%	ALGTRMI	6.5%
QQLAVDT	2.7%	GPLKTWK	6.8%	STLNWWN	2.7%
STLNWWN	1.4%	ESRVMSR	4.1%	KDNTYVM	1.0%
QTNMRAP	0.4%	GTSDKLP	2.3%	TTQVLEA	0.8%
ESRVMSR	0.3%	KDNTYVM	2.3%	GTSDKLP	0.5%
EARMHRA	0.3%	STLNWWN	2.1%	HSHTLTW	0.4%
TYPSLWK	0.3%	GTLASLN	1.3%	YPWFIRA	0.3%

CONTROL: C3p					
DVHKLGN	23.8%	DVHKLGN	37.4%	AQSLETS	40.5%
ISHTKAE	10.0%	AQSLETS	12.9%	ETALIAA	16.6%
NDIPHLM	9.1%	NDIPHLM	12.3%	QQLAVDT	11.0%
RIHVHNS	4.4%	ETALIAA	7.5%	DVHKLGN	10.8%
AQSLETS	3.5%	QQLAVDT	3.8%	LYADSVL	7.7%
SLTMKAN	2.1%	RIHVHNS	3.5%	RIHVHNS	3.1%
DLNMPIS	2.0%	LYADSVL	3.2%	NDIPHLM	2.8%
LYADSVL	1.9%	SVASTTV	3.2%	NPMYHSS	1.9%
ITPAYDN	1.8%	DLNMPIS	2.1%	KQEAMTL	1.1%
SVASTTV	1.7%	IAGPFKL	1.6%	QNSPFSE	1.0%

Supplementary Table S5. The ten most frequent peptides and their frequencies for each sample. The peptides are separated according selection round (Round 1-3). The peptides with a grey overlay were also observed in a control sample with a prevalence > 1.

Peptide	# of samples		Peptide	# of samples
QLYREFN	11		AERYPDS	3
SLLGQTP	8		DSSLFAL	3
TTQVLEA	7		DSSSFAL	3
VSRDTPQ	7		DTALHSL	3
IDRTQFM	7		EFARNSI	3
WSLSELH	7		GGPSGKL	3
YPWFIRA	7		GLAPFNA	3
DAIPTSV	6		GQSEKHL	3
KMISATE	6		GWETRME	3
NEAPRHA	6		IRVPPLV	3
AQYVAVG	5		KFAASSY	3
DGRYYIN	5		LSKINSP	3
DVHKLGN	5		LVIHSRT	3
ENHVHVR	5		LYADSVL	3
SNYHPSI	5		NEHTGNL	3
AQSLETS	5		NPPWFHT	3
DALAATL	5		NVRYNMI	3
DNYDSTA	5		QQLAVDT	3
GLRNPPS	5		QQTNWSL	3
IDNSHTH	5		SAAWNKS	3
TTANVRI	5		STPATLI	3
TNESYRH	4		TNLSHVP	3
TVISQNM	4		VPILNSL	3
EMIYEAN	4		VVTPKTA	3
QLSLIST	4		YYNTTPN	3
STIVAEM	4		ALGTRMI	3
DLNMPIS	4		EVMGRLA	3
GTGSQAS	4		GASESYL	3
GTSIYLH	4		GGSTDCI	3
HGGVRLY	4		GHDPTPL	3
HLNQQNH	4		LAKPLLV	3
HYIDFRW	4		LSANHWV	3
IHMPASP	4		MHTVAVQ	3
SIMNDRY	4		MPTWLHH	3
AYVARQN	4		SQPTWMF	3
KHLTAMA	4		SYTDLLR	3
SHVNVPS	4		THNLPVV	3
SSNQFHQ	4		TLSERNI	3
DSMSLLQ	3		TVTFIDS	3
HWSDDPH	3		VSGLLVE	3
SDRVLDN	3		YSLKQYQ	3

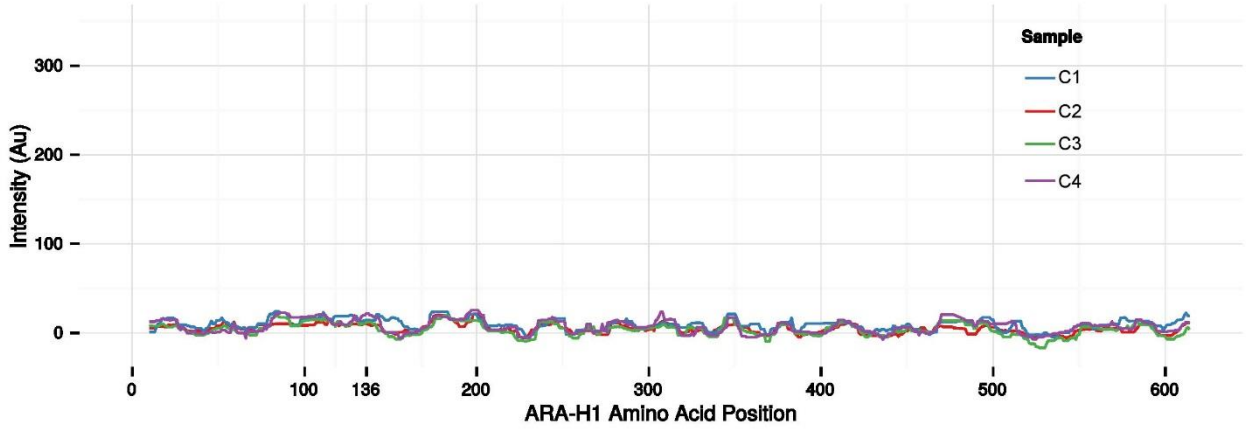
Supplementary Table S6. Dominating TUP candidates. Peptides that have been identified with a peptide prevalence >1 in a minimum of 3 different samples, including at least 1 control sample, are listed. The number of samples were a given peptide was observed is listed in the "# of samples" column.

Round 2	Round 3
IWRLPTH	IWRLPTH
WRLPYHN	HWRLPLH
WRHPSHF	WRLPYHN
WRLPYHN	WRHPSHF
TWRHPHH	WRHPSHF
HWRLPLH	TWRMPYH
TWRMPYH	WRLPYHN
QWQRPMH	TWRHPHH
WRLPYHP	VWRLPTH
WRLPEHR	QWQRPMH
IWRLPLH	WRLPYHN
LRLGPTH	TWKHPHH
WSLPELH	WWHPSHF
WRHPTHD	WRLPEHR
QWQRPMH	WRHPTHD
HWRLPTH	TWRMPYH
WLRLSTH	IWRMPTH
HWRLPLH	
HWQLPLH	
IWRPAAV	
AWRLPLH	
WRYPSHF	
WRMPLHR	
WRDPSHF	
IWRLPPQ	
WRHPSHF	
TWRMPYH	
IWRLQTH	

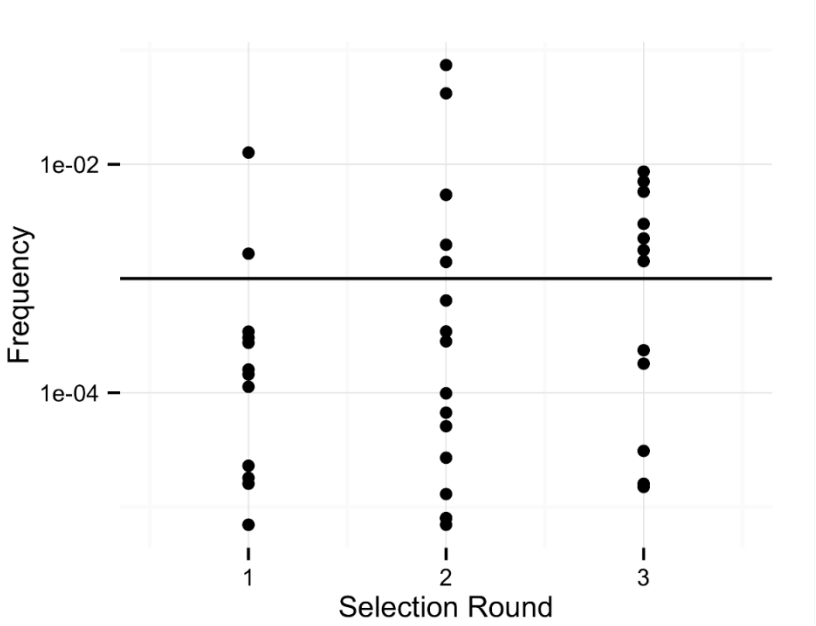
Supplementary Table S7. List of peptides in the significant peptide clusters separated according to selection round.

	Round 2	Round 3
P1	13	8
P1_12	4	3
P2	0	0
P3	4	4
P4	2	0
P4_12	2	1
Controls	3	1
Combined	28	17

Supplementary Table S8. The number of peptides in the significant clusters in selection round 2 and 3 separated according to the patient sample they were derived from.



Supplementary Figure S4. Intensity based on control IgE reactivity to Ara h 1 as measured by peptide micro-arrays. The right-bound rolling median (window size 12) of the mean intensity of the triplicate 12-mer peptides overlapping each residue is shown for every control sample. The start position of the epitope identified in the patients, at position 136, has been specifically marked.



Supplementary Figure S5. Frequency of peptides that were included in the significant cluster. The logarithmic peptide frequencies are separated according to selection round (1-3). The horizontal line represents the suggested detection limit of traditional phage display approaches, assuming that 1000 phage colonies are assayed thereby detecting peptides with a frequency above 0.001. Peptides with a rank score of 0 have not been plotted.

Appendix B

Supplementary information – Chapter 3

Here follows 6 tables of the 10 most frequent peptides in each of 6 analyzed animal sera. Numbers 1,2 and 6 refers different animals. A or B refers to baseline and after immunization, respectively.

1A					
Round 1		**Round 2**		**Round 3**	
SQPTWMF	3.2%	SGYQLEM	20.6%	SGYQLEM	23.4%
ADAKTVT	2.8%	SQPTWMF	7.4%	SQPTWMF	11.3%
FQLTYPT	2.7%	AGFITRD	6.3%	AGFITRD	9.0%
TSAHHRM	2.4%	TGFLVNV	2.9%	AGYVRHM	5.5%
VDFEEPN	1.7%	IEAQSMK	2.6%	TGFLVNV	4.1%
TGFLVNV	1.5%	TSAHHRM	1.9%	NTQLHPS	3.9%
NTLYNHR	1.5%	TGFHLTM	1.6%	TGFHLTM	3.9%
DIQEWSL	1.4%	YPLDSRV	1.6%	GQSEKHL	2.7%
NWDPVMR	1.4%	SGYLTNP	1.5%	VENVHVR	2.7%
SGYQLEM	1.2%	VENVHVR	1.4%	SGYLTNP	2.6%

1B					
Round 1		**Round 2**		**Round 3**	
NSIYQAW	10.2%	NSIYQAW	44.7%	NSIYQAW	45.3%
SSAVNVM	4.0%	YQWELYS	13.9%	YQWELYS	15.6%
YQWELYS	2.0%	SHASWQT	3.9%	GQSEKHL	8.4%
MKLPITL	1.8%	GTGSQAS	3.4%	SGYQLEM	5.9%
TARNLNF	1.5%	GQSEKHL	3.4%	LPVRLDW	3.3%
SHVTMNR	1.5%	SSAVNVM	2.7%	QTTFDRN	3.2%
SFAPSVS	1.3%	GLRNPPS	2.6%	QVYLVAA	2.1%
GTGSQAS	1.2%	WGNSRHM	2.4%	GMHETHV	2.0%
LPMPSHK	1.1%	GMHETHV	1.6%	GTGSQAS	1.6%
APKSGFT	1.0%	HLNQQNH	1.3%	AFPTMGN	1.5%

2A					
Round 1		Round 2		Round 3	
TVNFKLY	6.4%	TVNFKLY	32.6%	TVNFKLY	43.7%
QLAVAPS	2.8%	GPLHAQF	9.9%	GPLHAQF	16.0%
AAKTMSW	2.5%	AAKTMSW	9.3%	AAKTMSW	7.9%
ETALIAA	1.9%	QLAVAPS	4.8%	ALGTRMI	5.0%
TPYMTPK	1.6%	ETALIAA	3.1%	QLAVAPS	4.9%
KLTVNPV	1.6%	SQPTWMF	2.5%	HVPRAMA	4.2%
TNIVTGS	1.5%	HVPRAMA	2.3%	LSNPVMY	2.7%
IDRTQFM	1.3%	VPNPANY	1.7%	VPNPANY	2.6%
THNMHES	1.2%	TNIVTGS	1.7%	ETALIAA	1.6%
ALGTRMI	1.0%	SYASYYL	1.6%	DSSLFAL	1.5%

2B					
Round 1		Round 2		Round 3	
TMTVYTK	17.0%	TMTVYTK	17.5%	TVNFKLY	37.6%
SATWNPV	3.6%	TVNFKLY	14.9%	QLAVAPS	10.2%
SHENFTS	3.2%	QLAVAPS	8.2%	DSSLFAL	8.4%
YTNHQPQ	2.7%	DSSLFAL	6.4%	YGAKNNL	7.4%
QEVVTNH	2.0%	SATWNPV	5.2%	TMTVYTK	6.5%
QLAVAPS	1.4%	YTNHQPQ	4.5%	GPLHAQF	4.7%
TSAHHRM	1.3%	YGAKNNL	3.2%	SAHMRTM	2.3%
VSGFRID	1.2%	SHENFTS	3.0%	HYIDFRW	2.2%
TVNFKLY	1.1%	TSAHHRM	2.7%	ESRVMSR	2.0%
NTQLHPS	1.0%	GTSIYLH	2.4%	TSAHHRM	1.6%

6A					
Round 1		**Round 2**		**Round 3**	
GLLEIMG	7.4%	GASESYL	11.0%	TVNFKLY	29.2%
MPDLLRR	4.6%	TVNFKLY	9.9%	GASESYL	23.3%
NHITGYS	4.4%	AQQFPFR	8.8%	VENVHVR	5.4%
HAGFVPS	2.5%	SVPHSLN	7.4%	SVPHSLN	5.3%
NGSAPMT	1.9%	HAGFVPS	4.4%	AQQFPFR	4.9%
NASPSPT	1.8%	NHITGYS	4.1%	HGGVRLY	3.0%
TPKFMRY	1.8%	RHVEAHL	3.4%	NVKSAIT	2.4%
HTYELMR	1.8%	NVKSAIT	3.0%	HAGFVPS	2.4%
LPLTPPQ	1.8%	SIPPRNL	3.0%	GQSEKHL	2.0%
MPYWTGS	1.5%	QYVTPKW	2.6%	SQPTWMF	1.9%

6B					
Round 1		**Round 2**		**Round 3**	
YSEPAVT	5.2%	YSEPAVT	11.5%	YSEPAVT	14.8%
VPNIVTQ	3.6%	HSHTLTW	7.4%	HSHTLTW	11.9%
QIYSQKN	3.5%	SYTDLLR	6.7%	TVNFKLY	7.7%
APFNFGN	2.7%	VPNIVTQ	6.6%	VPNIVTQ	7.6%
SYTDLLR	2.4%	APFNFGN	5.2%	APFNFGN	6.6%
NTDFRAT	2.2%	LEPRMAN	3.5%	NQIYSAN	5.2%
GRVPYPV	1.7%	TVNFKLY	2.6%	GSLSRFI	3.2%
EMSTAMN	1.7%	KMISATE	2.4%	SYTDLLR	3.1%
VHAISSA	1.6%	SFSRAES	2.3%	GPSASRN	2.9%
SFSRAES	1.6%	TMQNIPN	2.3%	KMISATE	2.5%

Appendix C

Supplementary information – Chapter 4

Supplementary information – High-resolution epitope mapping of peanut allergens demonstrates individualized and highly persistent antibody binding patterns at the residue level

Christian S. Hansen[1], Martin Dufva[2]*, Katrine L. Bøgh[3], Eric Sullivan[4], Jigar Patel[4], Thomas Eiwegger[5,6], Zsolt Szépfalusi[5], Morten Nielsen[1,7], Anders Christiansen[2]

[1] Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark

[2] Department of Micro- and Nanotechnology, Technical University of Denmark, Kgs. Lyngby, Denmark,

[3] National Food Institute, Technical University of Denmark, Søborg, Denmark

[4] Roche NimbleGen, Madison, Wisconsin, the United States of America

[5] Department of Pediatrics, Medical University of Vienna, Vienna, Austria

[6] Department of Pediatrics, Hospital for Sick Children, University of Toronto, Toronto, Canada

[7] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

* Correspondence and requests for materials should be addressed to Martin Dufva (Dufva@nanotech.dtu.dk)

Patient ID	Sample ID	Age at sampling	Gender	Symptoms
P1	P1_06	18	f	Anaphylaxis; bronchial obstruction, drop blood pressure, angioedema, oral edema, eczema, uritcaria, gastro-intestinal symptoms
	P1_08	22		
	P1_12	26		
P2	P2_03	8	f	Anaphylaxis, bronchial obstruction, urticaria
	P2_07	12		
P3	P3_03	26	m	Anaphylaxis, drop of blood pressure; gastro-intestinal symptoms
	P3_05	28		
	P3_13	36		
P4	P4_04	30	f	Angioedema, gastro-intestinal symptoms
	P4_14	40		
P5	P5_03	22	m	breathlessness, bronchial obstruction, gastrointestinal symptoms
	P5_12	31		
P6	P6_03	25	m	Gastro-intestinal symptoms, oral edema, bronchial obstruction
	P6_08	30		
	P6_12	34		
P7	P7_06	33	m	Angioedema, urticaria
	P7_14	41		

Table E1: Patient sample overview and clinical symptoms.

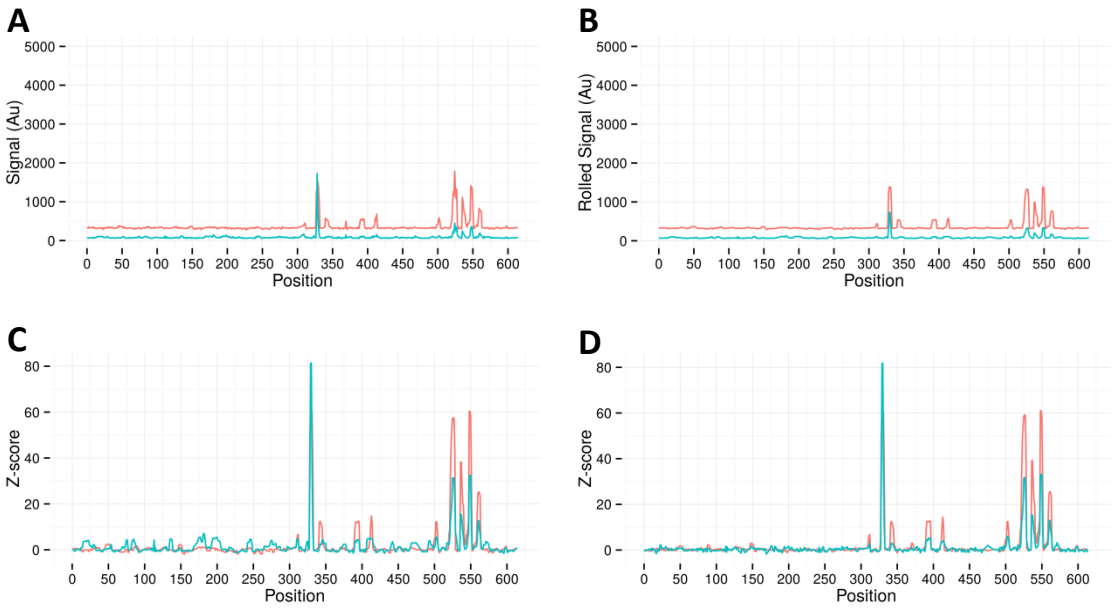


Figure E2. Illustration of signal normalization of IgE (red) and IgG4 (blue) signals of peptides probed with sample P5_12. **(A)**: Raw fluorescence signal (median of triplicates). **(B)**: Signal smoothing using a rolling median across x-axis with a window size of 3. **(C)**: Z-score of rolling median against baseline levels calculated from non-relevant peptides. **(D)**: Corrected Z-score: The Z-score of each peptide is subtracted the maximum of identical peptides probed with control samples.

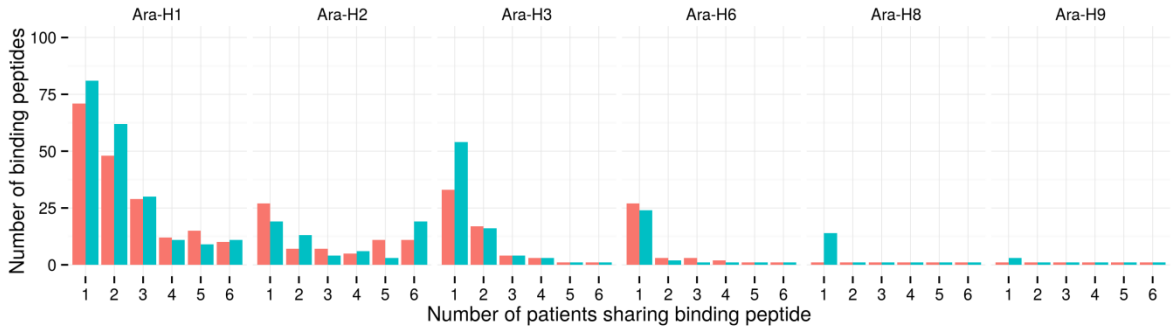


Figure E3. Patient overlap of IgE (red) or IgG4 (blue) binding peptides in all six allergens. Binding peptides are defined as native peptides having a corrected Z-score above 2 (see methods section for details).

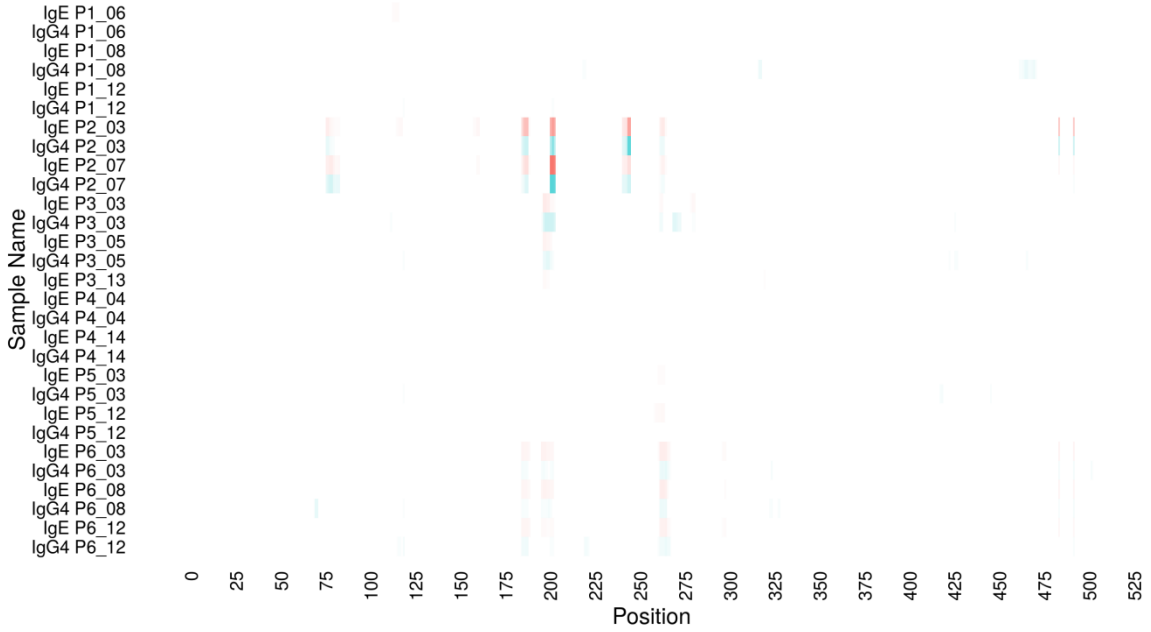


Figure E4. Corrected Z-score of native peptides originating from Ara h 3. The Z-score is reported at the starting position of a given peptide in the allergen sequence (x-axis) for both IgE (red) and IgG4 (blue). Each row indicates IgE or IgG4 channels from patient samples P1 through P6 at a certain year of sampling (e.g. P1_06 is Patient 1 sampled in 2006).

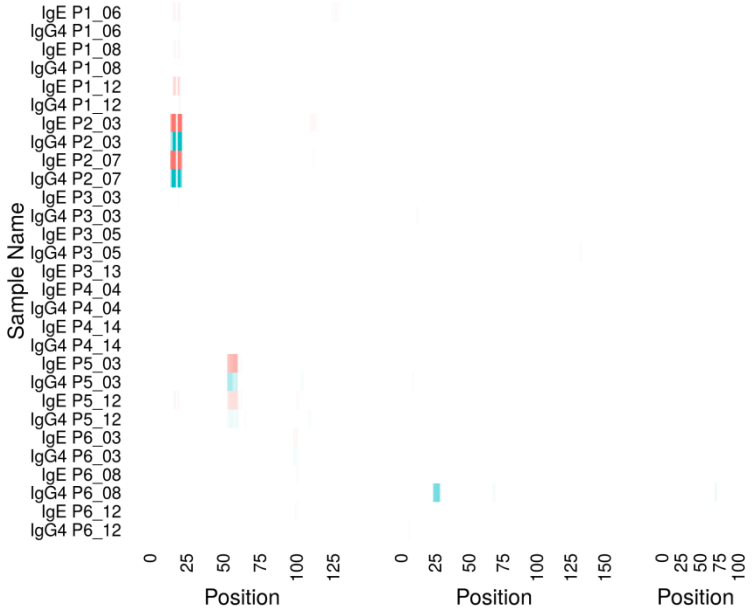


Figure E5. Similar to Figure E4, for Ara h 6 (left), Ara h 8 (middle) and Ara h 9 (right).

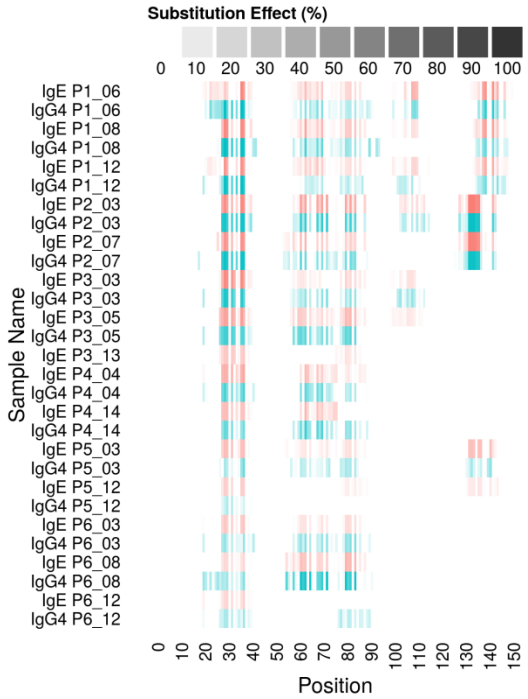


Figure E6. Alanine substitution effect in Ara h 2 for residue-level epitope mapping. The relative effect of substituting a given position to an alanine are reported as 0% to 100% loss of signal compared to the signal of the native 12-mer peptides. The substitution effects are reported as the median across multiple substitutions of the protein residue (x-axis) being expressed in overlapping binding peptides (see the Methods section for details). Each row indicates IgE or IgG4 channels from patient samples P1 through P6 at a certain year of sampling (e.g. P1_06 is Patient 1 sampled in 2006).

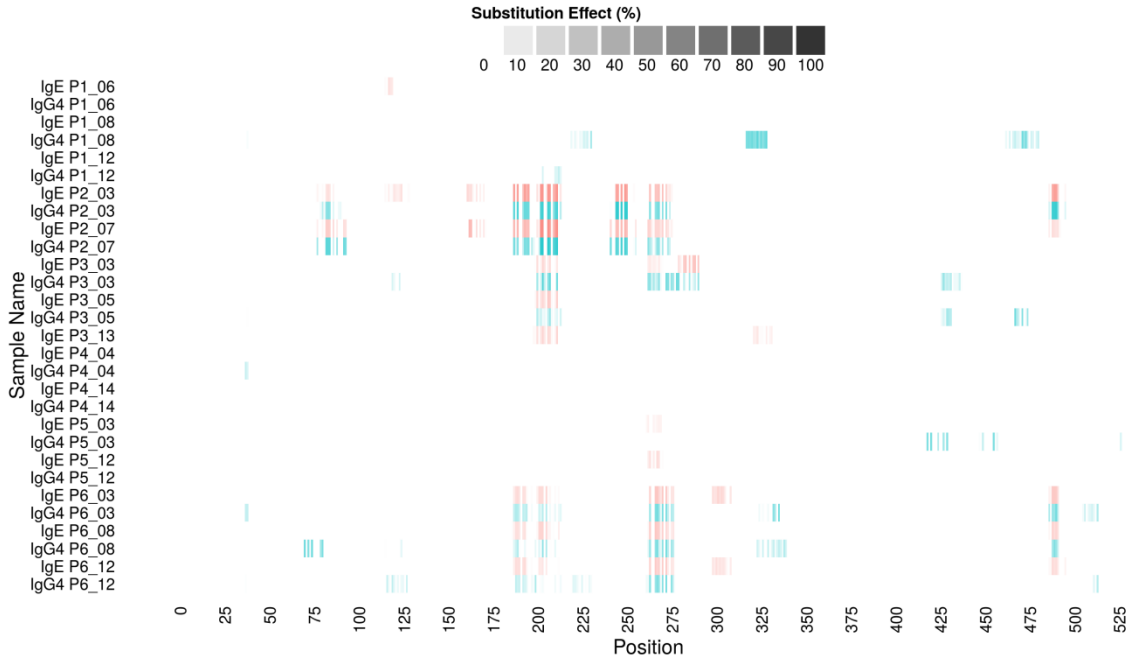


Figure E7. Similar to Figure E6, for Ara h 3.

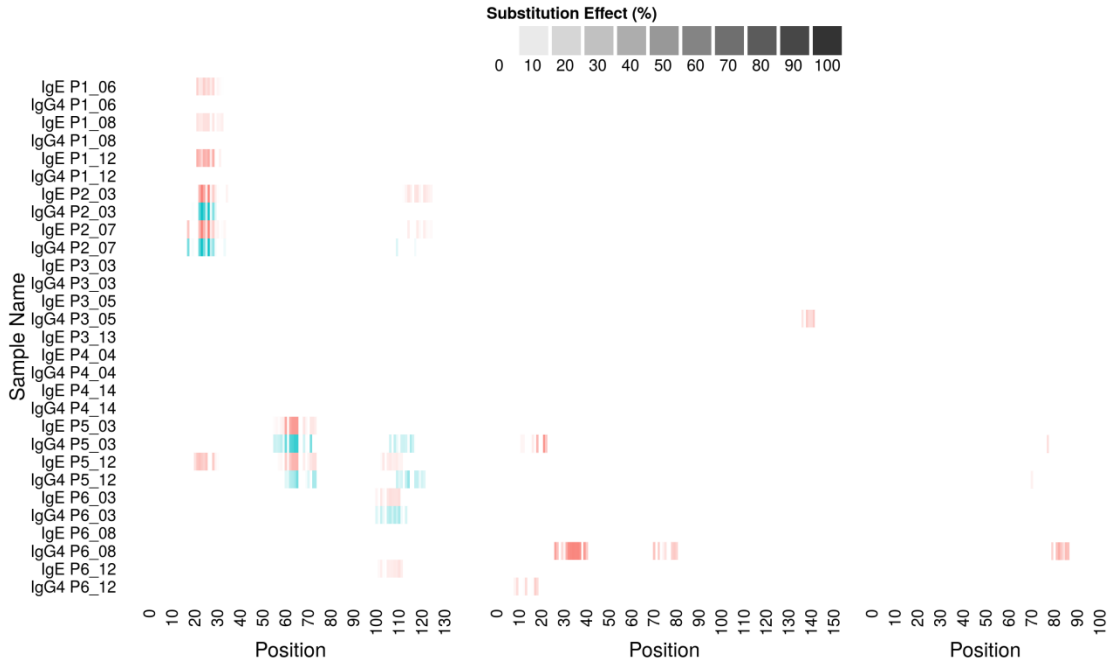


Figure E8. Similar to Figure E6, for Ara h 6 (left), Ara h 8 (middle) and Ara h 9 (right).

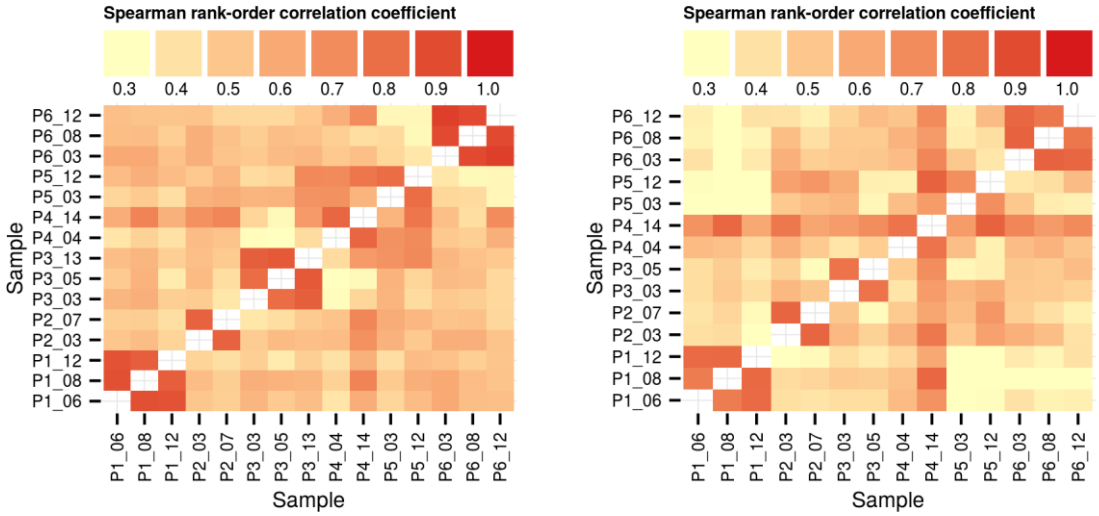


Figure E9. Correlation heatmaps of Spearman's rank correlation between substitution effects of allergen residues between samples. The results for IgE are shown to the left, and for IgG4 to the right. Only substitutions of positive peanut peptides in both samples are included.

A

		P1_06		P1_08	
		IgE	IgG4	IgE	IgG4
26	ELQGDRRCQSQL				
27	LQGDRRCQSQLE	QGDRR Q Q E	QGDRR Q Q E	QGDRR Q QLE	QGDRR Q QLE
28	QGDRRCQSQLER	RR ER	DRR Q Q ER	DRR Q ER	DRR Q Q ER
29	GDRRCQSQLERA	DRR ER	DRR Q Q ER	DRR Q ER	DRR Q Q ER
30	DRRCQSQLERAN	RR Q ER	DRR Q Q ER	DRR Q Q ER	DRR Q Q ER
31	RRCQSQLERANL	RR Q Q NL	RR Q Q N	RR Q QL NL	RR QSQL L
32	RCQSQLERANLR		Q Q R		Q R

		P2_03		P2_07	
		IgE	IgG4	IgE	IgG4
26	ELQGDRRCQSQL				
27	LQGDRRCQSQLE	DRR Q QLE	R	GDRR Q QLE	R Q QL
28	QGDRRCQSQLER	DRR Q QLER	DRR Q Q ER	Q DRR Q Q ER	Q DRR Q Q ER
29	GDRRCQSQLERA	DRR Q Q ER	DRR Q Q ER	DRR Q Q ER	DRR Q Q ER
30	DRRCQSQLERAN	DRR Q Q ER	DRR Q Q ER	DRR Q ER	DRR Q Q ER
31	RRCQSQLERANL	R Q Q NL	R CQ Q NL	R Q Q	R Q Q
32	RCQSQLERANLR				

		P3_03		P3_05	
		IgE	IgG4	IgE	IgG4
26	ELQGDRRCQSQL				
27	LQGDRRCQSQLE	LQGDRR QS LE	GDRR QS LE	LQGDRRCQS LE	GDRR QS E
28	QGDRRCQSQLER	DRR QS ER	DRR QS ER	GDRR QS LER	GDRR QS LER
29	GDRRCQSQLERA	GDRR QS ER	DRR QS ER	GDRR QS ER	GDRR QS ER
30	DRRCQSQLERAN	DRR QS ER	DRR QS ER	DRR QS LER N	DRR QS ER
31	RRCQSQLERANL	RR QS L NL	RR QS L	RR QS L N	RR QS N
32	RCQSQLERANLR				

B

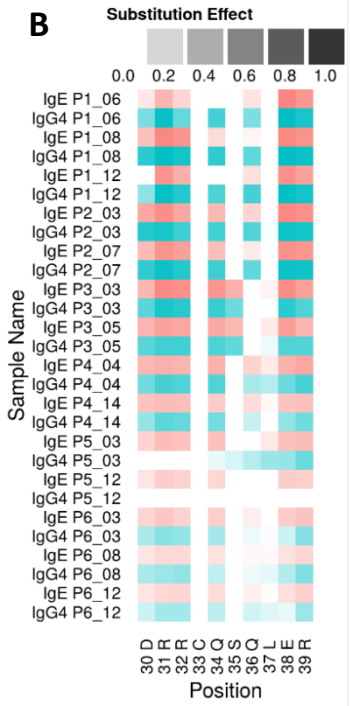
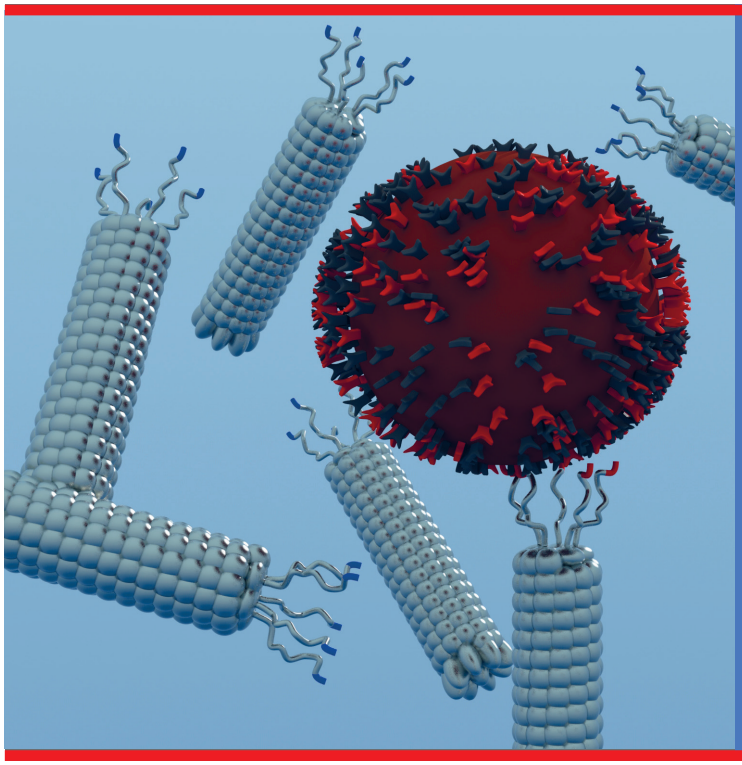


Figure E10: Alanine substitution effect for a sample epitope in Ara h 2 shown at high-resolution. IgE effect is shown in red and IgG4 effect in blue. **(A)** Alanine substitution analysis of 7 individual peptides spanning the 26-44 region of Ara h 2. Rows indicate each sliding window of overlapping 12-mer peptides undergoing substitution. Columns indicate patients P1, P2 and P3 as well as the year of sampling, e.g. P1_06 indicate Patient P6 sampled in 2006. Peptide residues are shown in a zero to full intensity scale representing a 0-100% decrease in signal upon alanine substitution, see methods section for details. **(B)** Median substitution effects of Ara h 2 protein residues 30-39 represented in multiple overlapping 12-mer peptides. See the Figure 3 legend for details.



Published by:
DTU Nanotech
Department of Micro- and Nanotechnology
Technical University of Denmark
Ørsteds Plads, building 345B
DK-2800 Kgs. Lyngby